UNIVERSITÉ AIX-MARSEILLE I - PROVENCE
U.F.R. de Mathématiques, Informatique et Mécanique
École Doctorale Mathématiques et Informatique de Marseille – E.D. 184

THÈSE

Pour obtenir le grade de
Docteur de l'université Aix-Marseille I
Discipline : Mathématiques

présentée et soutenue publiquement
par

Nicolas Klutchnikoff

le *14 décembre 2005*

Sur l'estimation adaptive de fonctions anisotropes

Directeur de thèse : **M. Oleg Lepski**

Rapporteurs :

Boris Levit	Professeur à Queen's University, Ontario
Dominique Picard	Professeur à l'Université de Paris VII

Jury :

Gérard Biau	Professeur à l'Université de Montpellier II
Yuri Golubev	DR à l'Université de Provence
Marc Hoffmann	Professeur à l'Université de Marne-La-Vallée
Oleg Lepski	Professeur à l'Université de Provence
Dominique Picard	Professeur à l'Université de Paris VII

Table des matières

Remerciements

Je tiens à remercier toutes les personnes qui, d'une manière ou d'une autre, m'ont permis de réaliser cette thèse dans les meilleures conditions.

En premier lieu, mes pensées vont à mon directeur de thèse, Oleg Lepski, qui m'a proposé un sujet extrêmement intéressant. Durant ces trois années, il n'a pas compté le temps consacré à mes travaux et ses conseils et remarques, toujours précis (même lorsque mes questions ne l'étaient pas !), m'ont souvent éclairés et motivés. Enfin, je crois qu'il a su me communiquer une partie de son enthousiasme pour cette branche des mathématiques si passionante.

Merci à Boris Levit et Dominique Picard d'avoir accepté, malgré un délai assez bref, de s'acquitter d'une charge de travail supplémentaire en rédigeant les indispensables rapports de thèse. Je suis également très reconnaissant à Gérard Biau, Youri Golubev et Marc Hoffman de participer au jury.

Lors de la préparation de cette thèse, j'ai eu la chance de faire parti de l'équipe de probabiltés et statistique au sein du LATP. Je souhaite remercier chaleureusement tous ses membres pour la bonne humeur générale dans laquelle j'ai été accueilli. Au moment de terminer cette thèse, je me dois de remercier plus personnellement Fabienne Castell qui, lorsqu'elle me préparait à l'agrégation, a su me donner envie de plonger dans l'univers de l'aléatoire...

J'ai eu beaucoup de plaisir à cohabiter avec les membres successifs du bureau R132. Karelle, Alexandre, Rémi et plus récemment Simona ont été d'une compagnie très agréable. Il semble d'ailleurs que ce soit une règle pour tous les membres du CMI que j'ai croisé et avec qui j'ai eu le plaisir de parler en prenant un café au foyer.

Je tiens également à remercier mes amis pour tous les bons moments passés ensemble. Je pense en particulier à Stéphanie, Florence, Léa, Julie, Muriel, Sébastien et Nicolas pour les matheux et à Cyrille, Laurence, Valérie, Na-

thalie, Carine, Olivier, Romain, Pierre, etc. pour les autres. J'espère que les « oubliés » ne m'en voudront pas trop...

Enfin, bien plus qu'un simple merci à mes parents, ma sœur, mes grand-mères...

... et à ma femme Sylvaine.

Première partie

Introduction

Chapitre 1

Introduction

1.1 Objectifs de la thèse

1.1.1 Présentation du modèle

Cette thèse présente des résultats d'estimation adaptative de fonctions (ou signaux) multidimensionnelles anisotropes, c'est-à-dire dont les régularités dans les différentes directions de l'espace diffèrent. Les résultats obtenus le sont pour des pertes ponctuelles dans le modèle de bruit blanc gaussien.

Plus précisément, nous cherchons à estimer des fonctions qui appartiennent à des classes d'espaces de Hölder anisotropes $H(\beta, L)$, où β est un vecteur représentant la régularité des fonctions et L une constante de Lipschitz.

Explicitons maintenant le modèle dans lequel se situent nos travaux. Nous observons une « trajectoire » d'un processus aléatoire X_ε satisfaisant l'équation différentielle stochastique suivante :

$$X_\varepsilon(du) = f(u)du + \varepsilon W(du), \quad u \in [0;1]^d, \tag{1.1}$$

où $f : \mathbf{R}^d \to \mathbf{R}$ est un signal inconnu à estimer, W est un bruit blanc gaussien standard de $\mathbf{R}^d$ dans $\mathbf{R}$ (voir par exemple []) et $\varepsilon > 0$ est un paramètre qui précise le niveau du bruit. Nous supposerons que ε est connu du statisticien.

De façon plus pratique, on peut observer toutes les quantités du type

$$\int_{\mathbf{R}^d} g(u) X_\varepsilon(du),$$

où g est une fonction connue de $\mathbf{L}^2(\mathbf{R}^d; \mathbf{R}) \triangleq \mathbf{L}^2$.

Remarque 1. *Le modèle de bruit blanc gaussien, à l'instard des lois gaussiennes, joue un rôle central en statistique puisqu'il approche d'autres modèles plus complexes, par exemple le modèle de densité non paramétrique. Citons par exemple les travaux de Nussbaum (1996).*

1.1.2 Différentes approches

Le but principal, lorsqu'on fait de l'estimation, est de choisir, parmis tous les estimateurs possibles fonctions mesurables des observations), le « meilleur ». Plusieurs approches et divers points de vue classiques s'offrent alors pour juger la qualité d'un estimateur.

1. L'approche minimax. On fixe un espace fonctionnel et un risque sur cet espace. Le meilleur estimateur est celui dont le risque est minimal. Cette approche suppose une connaissance *a priori* assez forte sur le signal à estimer puisqu'on suppose, par hypothèse, qu'il appartient à l'espace fonctionnel qu'on s'est fixé. Ce point de vue est donc parfois très limité en pratique.

2. L'approche adaptative au sens minimax peut être vue comme un développement de la première. On se fixe ici une famille d'espaces fonctionnels et un risque sur chacun d'entre-eux. Le meilleur estimateur est alors celui qui « minimise » simultanément tous les risques. La connaissance *a priori* sur le signal est donc nettement relâchée.

3. L'approche maxiset prend quant à elle le *contre-pied* de l'approche minimax. On se fixe un risque et une qualité d'estimation (vitesse) à atteindre. On calcule alors, pour chaque procédure, le plus grand espace fonctionnel — son maxiset — sur lequel elle atteint cette vitesse. Une procédure est d'autant meilleure que son maxiset est grand. En général on ne sait pas trouver de procédure ayant un maxiset maximal.

Dans cette thèse, nous nous intéressons particulièrement à la deuxième approche. Nous discuterons donc des approches minimax et adaptative au sens minimax dans la suite de cette introduction.

1.2 Théorie minimax

1.2.1 Généralités

On fixe, dans cette approche, un espace de Banach $\mathcal{F}$ (on suppose que le vrai signal appartient à cet espace) et une fonctionnelle $G : \mathcal{F} \to \Lambda$ où $(\Lambda, \|\cdot\|)$ est un second espace de Banach. Le but précis qu'on se fixe est d'estimer la fonctionnelle $G(f)$.

Étant donné un estimateur $\tilde{f}_\varepsilon$, on commence par mesurer sa qualité par rapport à chaque fonction $f \in \mathcal{F}$ par :

$$R_\varepsilon(\tilde{f}_\varepsilon, f) = \mathbf{E}_f\left[w(\|\tilde{f}_\varepsilon - G(f)\|)\right],$$

où $\mathbf{E}_f$ désigne l'espérance par rapport à la loi de $\mathcal{X}^{(\varepsilon)}$ (si f est la vrai valeur du signal inconnu) et où w est une fonction de perte à valeurs dans $\mathbf{R}^+$.

Comme on souhaite que l'estimateur qu'on choisira soit uniformément bon sur toutes les fonctions de $\mathcal{F}$, on introduit le risque maximal sur cet espace de la façon suivante :

$$R_\varepsilon(\tilde{f}_\varepsilon, \mathcal{F}) = \sup_{f\in\mathcal{F}} R_\varepsilon(\tilde{f}_\varepsilon, f).$$

L'idée étant de choisir comme estimateur celui dont le risque est minimal parmis tous les estimateurs possibles (on notera $\mathcal{E}$ cet ensemble), on introduit le risque minimax sur $\mathcal{F}$ défini par :

$$R_\varepsilon(\mathcal{F}) = \inf_{\tilde{f}_\varepsilon\in\mathcal{E}} R_\varepsilon(\tilde{f}_\varepsilon, \mathcal{F}).$$

Le but est alors double. On souhaite trouver à la fois un estimateur dont le risque est du même ordre, asymptotiquement (lorsque ε tend vers 0), que $R_\varepsilon(\mathcal{F})$, et d'autre part une formule explicite pour ce risque.

1.2.2 Méthode

La stratégie classique pour obtenir de tels résultats est la suivante. On commence à introduire pour toute normalisation (ou vitesse) $(\psi_\varepsilon)_{\varepsilon>0}$ le risque maximal renormalisé d'un estimateur $\tilde{f}_\varepsilon$:

$$R_\varepsilon(\tilde{f}_\varepsilon, \mathcal{F}, \psi_\varepsilon) = \sup_{f\in\mathcal{F}} \mathbf{E}_f\left[w(\psi_\varepsilon^{-1}\|\tilde{f}_\varepsilon - G(f)\|)\right].$$

On cherche ensuite à trouver une normalisation particulière φ_ε qui satisfait les deux inégalités suivantes :

Une borne supérieure qui assure que cette vitesse est atteignable (asymptotiquement) par un estimateur. C'est-à-dire qu'il existe un estimateur $\hat{f}_\varepsilon$ tel que :

$$\limsup_{\varepsilon \to 0} R_\varepsilon(\hat{f}_\varepsilon, \mathcal{F}, \varphi_\varepsilon) < +\infty.$$

Une borne inférieure qui assure quant à elle qu'on ne peut pas faire asymptotiquement mieux que φ_ε. C'est-à-dire :

$$\liminf_{\varepsilon \to 0} \inf_{\tilde{f}_\varepsilon \in \mathcal{E}} R_\varepsilon(\tilde{f}_\varepsilon, \mathcal{F}, \varphi_\varepsilon) > 0.$$

Si l'on trouve un tel couple $(\hat{f}_\varepsilon, \varphi_\varepsilon)$ on dit que c'est la solution du problème minimax sur $\mathcal{F}$. La vitesse φ_ε est appelée la vitesse minimax (asymptotique) et l'estimateur $\hat{f}_\varepsilon$ est dit (asymptotiquement) minimax.

Remarque 2. *On remarque que les deux inégalités ci-dessus impliquent que*

$$\varphi_\varepsilon \asymp R_\varepsilon(\mathcal{F}) \iff 0 < \liminf_{\varepsilon \to 0} \frac{\varphi_\varepsilon}{R_\varepsilon(\mathcal{F})} \leq \limsup_{\varepsilon \to 0} \frac{\varphi_\varepsilon}{R_\varepsilon(\mathcal{F})} < +\infty.$$

Dans cette thèse nous nous intéressons uniquement au cas où G est la fonctionnelle d'évaluation. C'est-à-dire qu'on fixe un point $t \in (0,1)^d$ et qu'on pose :

$$\begin{aligned} G : \mathcal{F} &\to \mathbf{R} \\ f &\mapsto f(t). \end{aligned}$$

Par ailleurs, $\|\cdot\|$ est remplacée par la valeur absolue dans $\mathbf{R}$ et nous choisirons de prendre comme fonction de perte $w_q(x) = x^q$ où $q > 0$ est fixé.

1.2.3 Résultats classiques

Rappelons quelques résultats classiques concernant l'estimation minimax dans le modèle de bruit blanc gaussien.

Tout d'abord la vitesse minimax sur les classes de Hölder unidimensionnelle $H(\beta, L)$ est connue pour l'estimation en norme $\mathbf{L}^p$ ($p \in [1; +\infty[$) et pour l'estimation ponctuelle où elle vaut $\varepsilon^{\frac{2\beta}{2\beta+1}}$ ainsi que pour l'estimation en norme $\mathbf{L}^\infty$ où elle vaut $(\varepsilon\sqrt{\ln 1/\varepsilon})^{\frac{2\beta}{2\beta+1}}$. On regardera par exemple Ibragimov et Hasminskii (1980,1981,1982) et Stone (1982).

Le cas multidimensionnel est plus complexe. Les premiers résultats ont été prouvés dans le cas isotrope (la régularité est la même dans chaque direction de l'espace). Stone (1980) a prouvé que, dans le modèle de régression, la vitesse d'estimation ponctuelle ou en norme $\mathbf{L}^p$ est donnée par $\varepsilon^{\frac{2\beta}{2\beta+d}}$ où d est la dimension. Un autre résultat donne un $\sqrt{\ln 1/\varepsilon}$ additionnel en norme $\mathbf{L}^\infty$. L'estimation est donc d'autant moins bonne que la dimension est grande.

Pour le cas anisotrope, Barron et al. (1999) ont exhibé la vitesse minimax pour l'estimation en norme $\mathbf{L}^2$. Elle dépend de la moyenne harmonique des régularités dans chaque direction de l'espace,

$$\frac{1}{\overline{\beta}} = \sum_{i=1}^{d} \frac{1}{\beta_i},$$

et est donnée par la formule $\varepsilon^{\frac{2\overline{\beta}}{2\overline{\beta}+1}}$. On ne voit plus ici la dépendence en la dimension de manière explicite : elle est cachée dans la formule de $\overline{\beta}$.

Enfin, dans Kerkyacharian et al. (2001), la vitesse minimax est exhibée sur des classes de Besov anisotropes. Les procédures construites pour atteindre ces vitesses ont été une source d'inspiration pour nos travaux.

D'autre part, certains résultats assez précis donnent même un équivalent exact de la vitesse minimax de convergence. Par exemple Bertin (2004) fournit un tel résultat pour l'estimation en norme $\mathbf{L}^\infty$ de fonctions höldériennes anisotropes en ajoutant la condition $\beta = (\beta_i)_{i=1,\dots,d} \in]0;1]^d$.

1.3 Théorie adaptative

1.3.1 Généralités

Pour l'approche adaptative, on ne suppose plus que le signal appartient à un espace fonctionnel connu exactement, ce qui peut être une gêne considérable en pratique, mais plutôt à une réunion de différents espaces fonctionnels. Cette réunion pouvant être très large, cette hypothèse est souvent beaucoup plus raisonnable. On note donc $(\Sigma(\varkappa))_{\varkappa\in\mathcal{J}}$ une famille d'espaces définis par un paramètre $\varkappa$ appelé « paramètre nuisible ».

En général, on suppose que $\mathcal{J} \subset \mathbf{R}^m$. On suppose également que le problème minimax est résolu sur chacun de ces espaces. On notera $N_\varepsilon(\varkappa)$ la vitesse minimax sur $\Sigma(\varkappa)$.

La stratégie est alors la suivante : on introduit un risque par espace fonctionnel. De façon formelle, à $\varkappa \in \mathcal{J}$ fixé, on considère le risque $R_\varepsilon^{(\varkappa)}(\cdot)$ défini pour tout estimateur $\tilde{f}_\varepsilon \in \mathcal{E}$ par :

$$R_\varepsilon^{(\varkappa)}(\tilde{f}_\varepsilon) = \sup_{f \in \Sigma(\varkappa)} \mathbf{E}_f \Big[w(\|\tilde{f}_\varepsilon - f\|) \Big] .$$

On souhaite alors trouver un seul estimateur dont la vitesse dépend de $\varkappa$ et tel que cet estimateur soit aussi précis que possible pour chaque valeur du paramètre nuisible.

1.3.2 Méthodes et résultats connus

Donnons un rapide aperçu des méthodes classiquement employées pour fabriquer des estimateurs adaptifs.

La sélection de modèles est une méthode particulièrement bien adaptée à l'estimation en norme $\mathbf{L}^2$. Elle utilise en effet assez fortement le caractère hilbertien (à travers la projection orthogonale) de cet espace.

La décomposition en série d'ondelettes et les techniques de seuillage permettent également d'obtenir des procédures adaptatives. Les travaux de Donoho et Johnstone (1994, 1995), Johnstone (1996) et Kerkyacharian et Picard (1997) jettent les bases de ces méthodes particulièrement performantes d'un point de vue pratique tout autant que théorique. L'idée de ces méthodes est assez simple, on décompose les observations dans une base d'ondelettes et l'on ne conserve que les coefficients qui dépassent un certain seuil. Le choix de ce seuil est un point crucial, et assez délicat, de cette théorie.

Les méthodes à noyaux sont basées quant à elles sur des idées de comparaison plus ou moins directes entre estimateurs à noyaux. L'idée est de construire une procédure qui choisit, en fonctions des observations, un estimateur à noyau parmi une collection d'estimateurs de ce type déterminée à l'avance. Ce choix se base en principe sur des considérations de type *comparaison biais-variance*. Citons par exemple les travaux de Lepski (1991), Tsybakov (2002) ou Bertin (2004) qui sont basées sur ce principe. Notons enfin que si l'on choisit un estimateur linéaire en fonction des observations, l'estimateur qui en résulte est non-linéaire. On sait par ailleurs les limites des procédures linéaires (cf. Nemirovski (1986), Donoho, Johnstone, Kerkyacharian et Picard (1996) ou encore Rivoirard (2004)).

Dans cette thèse, les méthodes développées sont des méthodes basées sur la comparaison d'estimateurs à noyau.

1.3.3 Le problème de l'optimalité

Dans la théorie adaptative, l'un des points délicats est de prouver que l'estimateur proposé est « optimal » en un certain sens. La première idée est de rechercher un estimateur qui minimise simultanément tous les risques $R_\varepsilon^{(\varkappa)}(\cdot)$. Il est facile de voir que pour réaliser cela il faut que cet estimateur atteigne la vitesse minimax sur chacun des espaces $\Sigma(\varkappa)$. S'il existe, un tel estimateur est alors appelé *adaptatif optimal*. Lepski (1991) a montré qu'il existe un estimateur adaptatif optimal (E.A.O) pour l'estimation en norme $\mathbf{L}^p$ ($p \in [2;\infty]$), sur des classes de Hölder unidimensionnelles $H(\beta)$ avec $\beta \in [\beta_*;\beta^*]$.

Malheureusement, il n'existe pas toujours d' E.A.O. Le premier résultat de cette nature est dû à Lepski (1991) qui le prouve pour l'estimation ponctuelle sur une famille d'espaces de Hölder unidimensionnels. Ceci se traduit par le fait que, pour tout estimateur $\tilde{f}_\varepsilon$ il existe un paramètre nuisible $\varkappa_0$ tel que :

$$\limsup_{\varepsilon\to 0} \sup_{f\in\Sigma(\varkappa_0)} \mathbf{E}_f\Big[w\left(N_\varepsilon^{-1}(\varkappa_0)\|\tilde{f}_\varepsilon - f\|\right)\Big] = +\infty.$$

Dans ce cas, il faut un critère qui permette de choisir le meilleur estimateur possible. Il existe actuellement plusieurs critères. Ils sont tous basés sur les idées suivantes :

1. Dans un premier temps, on définit des familles de normalisations dites *admissibles*. Une telle famille $\Psi^{(\varepsilon)} = \{\psi_\varepsilon(\varkappa)\}_{\varkappa\in\mathcal{J}}$ doit vérifier la propriété suivante : il existe un estimateur $\tilde{f}_\varepsilon$ tel que

$$\limsup_{\varepsilon\to 0} \sup_{\varkappa\in\mathcal{J}} \sup_{f\in\Sigma(\varkappa)} \mathbf{E}_f\Big[w\left(\psi_\varepsilon^{-1}(\varkappa)\|\tilde{f}_\varepsilon - f\|\right)\Big] < +\infty. \qquad \text{(B.S.A)}$$

   Autrement dit, $\tilde{f}_\varepsilon$ atteint simultanément sur chaque espace $\Sigma(\varkappa)$ la vitesse $\psi_\varepsilon(\varkappa)$.

2. Dans un deuxième temps, on propose un critère permettant de comparer les familles de normalisations admissibles pour choisir *la* « meilleure ». Notons que ce critère doit, en particulier, vérifier l'importante propriété suivante : si $N^{(\varepsilon)} = \{\varphi_\varepsilon(\varkappa)\}_{\varkappa\in\mathcal{J}}$ désigne la famille des vitesses minimax sur chaque espace de Hölder, et si cette famille est admissible, alors le critère proposé doit la chosir.

3. Enfin, on construit un estimateur f_ε^Φ qui vérifie (B.S.A) avec $\Psi^{(\varepsilon)} = \Phi^{(\varepsilon)}$. Un tel estimateur est alors appelé *estimateur adaptatif.*

La prochaine section est consacrée à l'étude de critères d'optimalité existant (Lepski (1991) et Tsybakov (1998)). Nous constaterons leurs défauts et proposerons un nouveau critère plus adapté à l'étude de procédures adaptatives en dimension supérieure à 1.

1.4 Optimalité de familles de normalisations

1.4.1 Constat préliminaire

Nous allons expliquer ici notre point de vue sur les critères d'optimalité pour choisir la « meilleure » famille de normalisations (vitesse adaptative) dans le cas où un estimateur adaptatif optimal n'existe pas, ce qui est le cas dans nos deux problèmes.

Nous revenons aux notations générales introduites à la section 1.3 pour une plus grande généralité.

Ces critères partent d'un même constat. Lorsqu'il n'existe pas d'E.A.O. il est toujours possible d'améliorer en un point (paramètre nuisible) au moins toute famille de normalisations admissible. En effet, notons $\Psi^{(\varepsilon)}$ une telle famille et $N^{(\varepsilon)}$ la famille des vitesses minimax sur chaque espace. Clairement, il existe au moins un paramètre nuisible $\varkappa_0$ où

$$\frac{\psi_\varepsilon(\varkappa_0)}{N_\varepsilon(\varkappa_0)} \xrightarrow[\varepsilon\to 0]{} +\infty,$$

sinon $N^{(\varepsilon)}$ serait admissible et alors il existerait un E.A.O par définition de l'admissibilité. Il suffit alors de choisir comme estimateur celui qui est minimax sur l'espace $\Sigma(\varkappa_0)$. Ailleurs, sa vitesse est éventuellement très mauvaise mais il améliore $\psi_\varepsilon(\varkappa_0)$.

Partant du constat qu'on ne peut pas empêcher une famille d'être améliorer il faut un critère qui garantit que le nombre de points où l'on peut améliorer la vitesse adaptative est **petit**.

1.4.2 Critères actuels

Commençons par rappeler les critères déjà existant. Il s'agit de comparer des familles de normalisations admissibles.

Lepski (1991). L'idée est d'introduire, pour chaque famille admissible $\Psi^{(\varepsilon)}$, la quantité suivante :

$$\Lambda_\varepsilon(\Psi) = \sup_{\varkappa \in \mathcal{J}} \frac{\psi_\varepsilon(\varkappa)}{N_\varepsilon(\varkappa)}.$$

Bien sûr, $\Lambda_\varepsilon(\Psi)$ tend vers $+\infty$ pour chaque $\Psi^{(\varepsilon)}$ admissible puisqu'on a supposé qu'il n'existe pas d'E.A.O.

On dit alors que $\Phi^{(\varepsilon)}$ est la vitesse adaptative si les deux conditions suivantes sont remplies :

1. $\Phi^{(\varepsilon)}$ minimise $\Lambda_\varepsilon(\cdot)$.
2. Si $\tilde{\Phi}^{(\varepsilon)}$ est une autre famille qui minimise $\Lambda_\varepsilon(\cdot)$, alors les deux familles sont, points à points, equivalentes en ordre, i.e.

$$\forall \varkappa \in \mathcal{J},\ \varphi_\varepsilon(\varkappa) \asymp \tilde{\varphi}_\varepsilon(\varkappa).$$

Même si cette définition est tout-à-fait convenable et fournit un premier critère intéressant, elle est trop globale. Expliquons-nous : les vitesses admissibles sont comparées par rapport à la vitesse de référence – qui est la vitesse minimax — et non pas entre elles. De plus, en considérant le maximum des rapports, on perd un renseignement précieux : le « nombre » de points où la vitesse est mauvaise (i.e. $\psi_\varepsilon(\varkappa)$ est grand). Par exemple, on pourrait imaginer une famille qui vaudrait, pour tout paramètre nuisible, la vitesse minimax, sauf en un point où elle serait très grande. Cette famille serait jugée mauvaise par ce critère alors qu'en pratique elle serait sans doute très bonne !

Le défaut majeur, à nos yeux, de ce critère peut donc être formulé de la manière suivante : une famille de vitesses trop lente pour une seule valeur du paramètre nuisible ne peut pas être la vitesse adaptative.

Tsybakov (1998). L'idée, dans ce critère, est de comparer les vitesses admissibles de manière ponctuelle. La vitesse adaptative va être comparée à chaque autre vitesse admissible et cela en chaque valeur du paramètre nuisible. Donnons le critère avant d'en expliquer la philosophie. La meilleure famille de normalisation $\Phi^{(\varepsilon)}$ (i.e. la vitesse adaptative) doit vérifier la propriété suivante : si $\Psi^{(\varepsilon)}$ est une autre famille admissible et s'il existe $\varkappa_0 \in \mathcal{J}$ tel que :

$$\frac{\psi_\varepsilon(\varkappa_0)}{\varphi_\varepsilon(\varkappa_0)} \xrightarrow[\varepsilon \to 0]{} 0,$$

alors il existe un autre point $\varkappa_1$ tel que :

$$\frac{\psi_\varepsilon(\varkappa_0)}{\varphi_\varepsilon(\varkappa_0)}\frac{\psi_\varepsilon(\varkappa_1)}{\varphi_\varepsilon(\varkappa_1)} \xrightarrow[\varepsilon\to 0]{} +\infty.$$

En d'autres termes, si l'on peut améliorer la vitesse adaptative en un point ($\varkappa_0$), cela se fait en perdant bien plus que ce que l'on a gané en un autre point ($\varkappa_1$).
Encore une fois, cette définition est valable pour tout problème d'estimation adaptative. Cependant, elle est un peu trop ponctuelle , au contraire de la précédente ! Par exemple, il peut y avoir, d'après la définition (ceci n'est pas le cas dans Tsybakov (2001)), plusieurs points $\varkappa_0$ et un seul point $\varkappa_1$. Ceci ne correspond pas nécessairement à l'idée qu'on se fait d'une vitesse optimale... De plus, l'unicité (à l'ordre près) des vitesses adaptatives n'est garantie que pour un nombre fini de paramètres nuisibles. Dans le cas d'une infinité, ce n'est plus évident.

1.4.3 Explications

Le dernier problème evoqué provient essentiellement du fait que cette définition est plus adaptée à la dimension 1 qu'au cas multidimensionnel. Expliquons-nous. On sait qu'on peut toujours améliorer une vitesse admissible en un point au moins et, en dimension 1, on ne sait pas faire beaucoup mieux dans la plupart des cas. Mais, nos résultats mettent à jour un tout autre phénomène en dimension $d \geq 2$. Effectivement, on prouve qu'on peut améliorer la vitesse adaptative, non seulement sur un espace, mais également sur une réunion d'espaces. Cependant, cette réunion est « petite » par rapport à la classe d'espaces de Hölder considérée.

Le nouveau critère que nous proposons tente de prendre en compte ces aspects. En particulier, il est à la fois moins global que le critère formulé par Lepski, et moins ponctuel que celui ennoncé par Tsybakov.

1.4.4 Un nouveau critère

L'idée est là encore de comparer les vitesses admissibles deux à deux. Pour avoir un aspect semi-global, on considère, pour deux vitesses $\Psi^{(\varepsilon)}$ et $\tilde{\Psi}^{(\varepsilon)}$ les sous ensembles suivants de $\mathcal{J}$:

$$\mathcal{I}_0\left(\Psi^{(\varepsilon)}/\tilde{\Psi}^{(\varepsilon)}\right) = \left\{\varkappa \in \mathcal{J}; \frac{\psi_\varepsilon(\varkappa)}{\tilde{\psi}_\varepsilon(\varkappa)} \xrightarrow[\varepsilon\to 0]{} 0\right\},$$

qui correspond à l'ensemble des points où $\Psi^{(\varepsilon)}$ est meilleure (donc plus petite) que $\tilde{\Psi}^{(\varepsilon)}$ et

$$\mathcal{I}_\infty\left(\Psi^{(\varepsilon)}/\tilde{\Psi}^{(\varepsilon)}\right) = \left\{\sigma \in \mathcal{J}; \frac{\psi_\varepsilon(\varkappa)}{\tilde{\psi}_\varepsilon(\varkappa)} \times \frac{\psi_\varepsilon(\sigma)}{\tilde{\psi}_\varepsilon(\sigma)} \xrightarrow[\varepsilon\to 0]{} \infty, \forall \varkappa \in \mathcal{I}_0\left(\Psi^{(\varepsilon)}/\tilde{\Psi}^{(\varepsilon)}\right)\right\},$$

qui représente, au contraire, l'ensemble des points où $\tilde{\Psi}^{(\varepsilon)}$ est meilleure que $\Psi^{(\varepsilon)}$. Remarquons qu'on impose même qu'elle soit bien meilleure à cause de la normalisation supplémentaire inspirée du critère de Tsybakov.

On peut alors donner une définition raisonnable de vitesse adaptative de la manière suivante.

Définition 1. *Une famille de normalisations $\Phi^{(\varepsilon)}$ est appelée une vitesse adaptative si elle vérifie la propriété suivante : pour toute famille admissible $\Psi^{(\varepsilon)}$, l'ensemble $\mathcal{I}_0(\Psi^{(\varepsilon)}/\Phi^{(\varepsilon)})$ est contenu dans une variété de dimension au plus égale à $m-1$ et, $\mathcal{I}_\infty(\Psi^{(\varepsilon)}/\Phi^{(\varepsilon)})$ contient un ensemble ouvert de $\mathbf{R}^m$.*

Exprimée différement, cette définition assure qu'une vitesse adaptative ne peut être améliorée que sur ensemble extrêmement réduit et qu'alors on perd bien plus, en faisant cela, sur un ensemble très massif.

De plus, il découle de la définition même de vitesse adaptative son unicité en ordre. C'est-à-dire que si $\Phi^{(\varepsilon)}$ et $\tilde{\Phi}^{(\varepsilon)}$ sont deux vitesses adaptatives, alors,

$$\forall \varkappa \in \mathcal{J},\ \varphi_\varepsilon(\varkappa) \asymp \tilde{\varphi}_\varepsilon(\varkappa).$$

En effet, raisonnons par l'absurde et supposons qu'il existe $\varkappa_0 \in \mathcal{I}_0(\Phi^{(\varepsilon)}/\tilde{\Phi}^{(\varepsilon)})$, alors $\mathcal{I}_\infty(\Phi^{(\varepsilon)}/\tilde{\Phi}^{(\varepsilon)})$ contient un ouvert car $\tilde{\Phi}^{(\varepsilon)}$ est une vitesse adaptative. Mais d'autre part, il est aisé de montrer que $\mathcal{I}_\infty(\Phi^{(\varepsilon)}/\tilde{\Phi}^{(\varepsilon)}) \subset \mathcal{I}_0(\tilde{\Phi}^{(\varepsilon)}/\Phi^{(\varepsilon)})$. Ceci est en contradiction avec le fait que $\Phi^{(\varepsilon)}$ est une vitesse adaptative, car $\mathcal{I}_0(\tilde{\Phi}^{(\varepsilon)}/\Phi^{(\varepsilon)})$ devrait être contenu dans un variété de dimension strictement inférieure à m.

Finalement, le critère que nous proposons est un raffinement des critères précédents qui offre le double avantage d'assurer l'uncité de la vitesse adaptative ainsi que d'être bien adapté au cas multidimmensionnel.

1.5 Les procédures

1.5.1 Généralités

L'optimalité d'une vitesse est une chose, mais encore faut-il construire un estimateur qui atteigne la vitesse adaptative. Dans cette thèse, nous construisons deux procédures adaptatives différentes : une pour résoudre le problème de l'adaptation totale, l'autre pour celui de l'adaptation partielle.

Ces deux procédures sont proches dans leur conception. Elles reposent en effet sur les idées apparues dans Lepski (1990) et développées pour le cas multidimensionnel dans Kerkyacharian et al. (2001) — même s'il s'agissait de construire une procédure *minimax* sur des espaces de Besov anisotropes.

Nous allons discuter des principes généraux qui nous ont guidés dans la construction des procédures qui seront exposées plus loin. Nous reviendrons alors plus précisément sur la spécificité de chaqu'une d'entre elles.

1.5.2 Principes

Le principe de construction des procédures basées sur le choix aléatoire d'un estimateur à noyau est assez simple et peut être expliqué formellement de la façon suivante. Si $\{\Sigma(\varkappa)\}_{\varkappa\in\mathcal{J}}$ est une famille d'espaces fonctionnels et si $\{\varphi_\varepsilon(\varkappa)\}_{\varkappa\in\mathcal{J}}$ est la vitesse adaptative qu'on souhaite atteindre, la méthode générale consiste à faire les choses suivantes.

1. Construire, pour chaque valeur $\varkappa$ du paramètre nuisible, un estimateur à nouyau qui atteint la vitesse $\varphi_\varepsilon(\varkappa)$ sur l'espace $\Sigma(\varkappa)$.
2. Choisir un nombre fini d'estimateurs assez grand pour bien estimer sur chaque espace $\Sigma(\varkappa)$: par exemple en faisant une grille suffisament fine de l'espace $\mathcal{J}$.
3. Introduire un système de comparaison des estimateurs entre eux pour en choisir un — en fonction des observations et donc de manière aléatoire — afin de l'utiliser pour l'estimation.

En règle générale, les points 1 et 2 ne posent pas les plus grandes difficultés même si quelques problèmes techniques (comme ce sera notre cas pour le deuxième point) peuvent se poser. En revanche, le troisième point est crucial.

Par exemple, on verra qu'on ne peut pas traité le cas multidimensionnel comme l'a été le cas réel dans Lepski (1991). En effet, dans ce dernier cas,

on choisit un estimateur en vérifiant une propriété assez simple (du type comparaison biais-variance), alors que nous serons obligés de comparer les estimateurs deux à deux à l'aide d'un critère plus élaboré et inspiré de Kerkyacharian et al. (2001).

1.5.3 Adaptation partielle

Ici, les deux premiers points sont calqués sur le cas de l'adaptation totale. Quelques modifications mineures interviennent (car on a des renseignements supplémentaires liés à l'information additionnelle dont on dispose), mais le cœur du problème est ailleurs.

La construction d'un esemble aléatoires de « bons » indices, $\mathcal{A}$, reste semblable au cas précédent, de même que le choix, par la procédure, d'un indice dans cet ensemble, s'il n'est pas vide.

Mais, dans le cas où $\mathcal{A}$ est vide, l'attitude adoptée est radicalement différente de ce que l'on faisait dans le cas de l'adaptation totale : estimer par n'importe quoi (dans notre cas précis, 0). En effet, le fait de n'avoir qu'un $\sqrt{\ln \ln 1/\varepsilon}$ dans la vitesse, au lieu de $\sqrt{\ln 1/\varepsilon}$, empêche d'avoir $\mathbf{P}[\mathcal{A}]$ très petite. Pour compenser cet état de fait, il faut pouvoir estimer avec une vitesse meilleure que ne le fait l'estimateur trivial.

Or, cette vitesse est exactement $(\varepsilon\sqrt{\ln 1/\varepsilon})^{2\gamma/(2\gamma+1)}$. Et donc, l'estimateur construit lors du cas précédent convient parfaitement.

Au final, la procédure qui atteint la vitesse optimale dans le cas de l'adaptation partielle dépend de la procédure construite pour résoudre le problème de l'adaptation totale.

1.6 Présentation des résultats

Nous allons maintenant expliquer plus en détail les principaux résultats obtenus durant cette thèse. Rappelons que nous nous sommes placés dans le modèle de bruit blanc et que nos observations $\mathcal{X}^{(\varepsilon)} = \{X_\varepsilon(u)\}_{u\in[0;1]^d}$ safisfont l'EDS (1.1).

Commençons par préciser les espaces fonctionnels ambiants.

1.6.1 Les espaces de Hölder anisotropes

Les fonctions qui appartiennent à ces espaces ont des régularités différentes suivant les axes définis par les vecteurs de la base canonique de $\mathbf{R}^d$. Il est donc naturel de définir, pour tout fonction $f : \mathbf{R}^d \to \mathbf{R}$ les fonctions auxiliaires suivantes

$$\forall y \in \mathbf{R}, \quad f_i(y|x) = f(x_1, \dots, x_{i-1}, x_i + y, x_{i+1}, x_d)$$

qui décrivent les accroissements de f au voisinage de tout point $x \in \mathbf{R}^d$, dans chacune des directions de base.

À partir de là, nous pouvons définir les espaces de Hölder anisotropes $H(\beta, L)$, pour tout vecteur $\beta = (\beta_1, \dots, \beta_d)$ tel que tous les β_i sont strictement positifs et pour tout $L > 0$.

Définition 2. *Une fonction $f : \mathbf{R}^d \to \mathbf{R}$ continue, à support compact inclus dans $[0;1]^d$ appartient à $H(\beta, L)$ si, et seulement si, les deux propriétés suivantes sont vérifiées :*

$$\sup_{i=1,\dots,n} \sup_{x \in \mathbf{R}^d} \sum_{s=0}^{m_i} \left\| f_i^{(s)}(\cdot|x) \right\|_\infty \leq L,$$

et, pour tout i et pour tout $y \in \mathbf{R}$,

$$\sup_{x \in \mathbf{R}^d} \left| f_i^{(m_i)}(y|x) - f_i^{(m_i)}(0|x) \right| \leq L|y|^{\alpha_i}$$

où m_i désigne le plus grand entier strictement inférieur à β_i et $\alpha_i = \beta_i - m_i$. En d'autre termes, dans la i^e direction de l'espace, f à la régularité höldérienne (β_i, L) au sens unidimensionnel.

Dans la suite nous ne considérons pas tous les espaces de Hölder mais seulement une sous famille. Fixons un vecteur $b = (b_1, \dots, b_d) \in (\mathbf{R}_+^*)^d$ et deux réels $l_* < l^*$. Nous notons

$$\mathcal{B} = \prod_{i=1}^{d}]0; b_i] \text{ et } \mathcal{I} = [l_*; l^*].$$

Nous travaillerons avec la famille suivante $\{H(\beta, L)\}_{(\beta,L) \in \mathcal{B} \times \mathcal{I}}$.

1.6.2 Nos objectifs

Nous nous sommes intéressés à deux problèmes d'estimation adaptative particuliers :

- le premier se situe dans la continuité des travaux classiques. Le but est de construire une procédure adaptative sur l'échelle de tous les espaces de Hölder considérés. Le résultat que nous obtenons généralise celui obtenu par Lespki (1991) en dimension 1. Nous nomerons ce problème « adaptation totale » ;
- le deuxième est plus original. Nous supposons que nous disposons d'une information supplémentaire sur le paramètre nuisible (ici le couple (β, L)). L'information considérée est du type $\bar{\beta}$ est connu. Nous expliquerons plus loin ce qui nous a fait considérer cette information suppmémentaire précise. Le résultat que nous obtenons est alors plus original en terme de vitesse obtenue. Ce second problème sera qualifié d'« adaptation partielle » ou « adaptation sous contrainte ».

Dans les deux cas nous prouvons l'optimalité de nos méthodes de différentes manières et en différents sens.

1.6.3 Adaptation totale

Commençons par exposer les résultats obtenus.

Théorème 1. *Il n'existe pas d'estimateur adaptatif optimal pour ce problème. En d'autres termes, pour tout estimateur $\tilde{f}_\varepsilon$, il existe $(\beta_0, L_0) \in \mathcal{B} \times \mathcal{I}$ tel que :*

$$\limsup_{\varepsilon \to 0} \sup_{f \in H(\beta_0, L_0)} \mathbf{E}_f \left[\left(N_\varepsilon^{-1}(\beta_0, L_0) |\tilde{f}_\varepsilon(t) - f(t)| \right)^q \right] = +\infty.$$

Rappelons que la vitesse minimax $N_\varepsilon(\beta, L)$ sur chaque espace $H(\beta, L)$ est connue et vaut :

$$L^{\frac{1}{2\beta+1}} \varepsilon^{\frac{2\bar{\beta}}{2\bar{\beta}+1}}.$$

Ce résultat n'est pas obtenu directement sous cette forme mais c'est un corollaire immédiat d'un résultat que nous obtenons.

Donnons tout de suite le résultat principal.

Théorème 2. *La famille de normalisation* $\Phi^{(\varepsilon)} = \{\varphi_\varepsilon(\beta, L)\}_{(\beta,L)\in\mathcal{B}\times\mathcal{I}}$ *définie par la formule :*

$$\varphi_\varepsilon(\beta, L) = L^{1/(2\bar{\beta}+1)} \left(\|K\|\varepsilon\rho_\varepsilon(\beta, L)\right)^{2\bar{\beta}/(2\bar{\beta}+1)}$$

où $\rho_\varepsilon(\beta, L)$ *est un terme de perte par rapport à la vitesse minimax qui vaut*

$$\rho_\varepsilon(\beta, L) = \sqrt{1 + \frac{4(\bar{b} - \bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)} \ln\frac{L}{\|K\|\varepsilon} + \frac{2}{2\bar{b}+1}\ln\frac{L}{l_*}}$$

est la ***vitesse adaptative*** *du modèle. De plus on construit un estimateur* f_ε^Φ *qui vérifie B.S.A avec* $\Psi^{(\varepsilon)} = \Phi^{(\varepsilon)}$.

Remarque 3. *Le terme de* vitesse adaptative *est à comprendre, ici, au sens de la nouvelle définition que nous avons donnée. En effet, ce critère a été conçu, en premier lieu, pour être appliqué au cas de l'adaptation totale. En effet, il n'est pas raisonnable dans ce cas d'utiliser un critère minimax (comme nous le ferons dans le cas de l'adaptation partielle) pour choisir une famille de normalisations optimale. Effectivement, de manière générale, la vitesse minimax sur une réunion ne peut pas être meilleure que la pire vitesse minimax sur chacun des espaces composant cette réunion. Dans notre cas c'est catastrophique lorsque* β *tend vers* 0 *!*

Par ailleurs, nous expliquons à la section suivante les idées principales de la construction de l'estimateur f_ε^Φ*. Il faut simplement remarquer pour l'instant que ce théorème assure que* f_ε^Φ *est un estimateur adaptatif. Le problème est donc résolu entièrement par ce théorème.*

Faisons quelques commentaires sur la vitesse adaptative trouvée. D'abord, constatons que la perte ρ_ε vaut 1 si $\bar{\beta} = \bar{b}$ (ce qui signifie simplement que $\beta = b$) et $L = l_*$. Donc en ce point, la vitesse obtenue est exactement la vitesse minimax sur $H(b, l_*)$. D'autre part ce point correspond également à l'espace de Hölder le plus régulier dans notre collection.

En dehors, la perte est de l'ordre de $\sqrt{\ln 1/\varepsilon}$ ce qui correspond exactement au cas classique unidimensionnel. Ce résultat généralise donc en tout point celui obtenu par Lepski (1990).

Notons maintenant un interprétation à propos de la perte ρ_ε. On a la formule suivante :

$$\rho_\varepsilon(\beta, L) = \sqrt{1 + 2\ln\frac{N_\varepsilon(\beta, L)}{N_\varepsilon(b, l_*)}}.$$

Donc, on paie exactement pour le rapport des vitesses minimax entre la vitesse de l'espace dans lequel se trouve le signal à estimer et la vitesse de l'espace le plus régulier de la collection.

1.6.4 Présentation de la procédure

Nous décrivons ici les points clés de la construction de l'estimateur que nous proposons. Signalons tout d'abord qu'un noyau vérifiant de bonnes conditions (en particulier d'orthogonalité par rapport à des polynômes) doit être fixé.

Le premier point exposé au paragraphe 1.5.2 ne pose aucun problème. Pour tout $(\beta, L) \in \mathcal{B}\times\mathcal{I}$, on construit une fenêtre $h(\beta, L, \varepsilon)$ qui permet d'atteindre, en utilisant l'estimateur construit sur cette fenêtre, la vitesse $\varphi_\varepsilon(\beta, L)$ sur $H(\beta, L)$.

Le deuxième point pose quelques petits problèmes techniques. Nous avons choisi de faire une grille dans l'espace des fenêtres plutôt que dans l'espace des régularités. Nous avons donc défini pour tout k dans $\mathbf{Z}^d$ des fenêtres $h^{(k)}$ et construit des indices $k(\beta, L, \varepsilon)$ tels que, à (β, L) fixé, l'estimateur construit en utilisant la fenêtre $h^{(k(\beta,L,\varepsilon))}$ est aussi bon (asymptotiquement) que celui utilisant la fenêtre $h(\beta, L, \varepsilon)$. Ceci est très comparable a ce qui a été fait dans Kerkyacharian et al. (2001). Le point délicat a été de trouver un sous ensemble de $\mathbf{Z}^d$, contenant tous les $k(\beta, L, \varepsilon)$, qui ne soit pas trop gros.

Ce point technique est particulier à notre problème. Il n'apparaissait pas dans l'article sus-cité. En effet, tous les indices considérés dans ce papier appartiennent à $\mathbf{N}^d$ qui est suffisament petit. Nous avons pu montrer que la famille $\{k(\beta, L, \varepsilon)\}_{(\beta,L)}$ était incluse dans une sorte de cône autour de $\mathbf{N}^d$. Ceci a été suffisant.

Pour le dernier point, la comparaison des estimateurs se fait deux à deux en introduisant, pour chaque comparaison, un estimateur artificiel qui permet un contrôle plus fin des biais. Cette technique est totalement différente de celle proposée dans le cas unidimensionnel par Lepski (1990) même si elle est basée sur une comparaison de type biais-variance.

1.6.5 Adaptation partielle

Commençons par expliquer plus clairement le problème. On considère un réel $0 < \gamma < \bar{b}$ et l'ensemble suivant :

$$\mathcal{B}(\gamma) = \left\{\beta \in \mathcal{B}; \bar{\beta} = \gamma\right\}.$$

Cet ensemble est constitué de paramètres de régularité qui donne, pour les espaces de Hölder associés à ces paramètres, des vitesses minimax égales entre elles. Il est donc aussi « facile » d'estimer sur chacun des espaces de la collection $\{H(\beta, L)\}_{(\beta,L)\in\mathcal{B}(\gamma)\times\mathcal{I}}$.

Remarque 4. *Le cas $\gamma = \bar{b}$ a été exclu car il n'est pas intéressant. En effet $\mathcal{B}(\gamma)$ est alors réduit à au seul point $\{b\}$.*

Une question naturelle est de se demander si la connaissance *a priori* du paramètre γ aide à trouver une procedure adaptative plus performante que f_ε^Φ. On peut même se demander s'il existe un estimateur adaptatif optimal.

Un premier résultat fournit une réponse négative à cette question.

Théorème 3. *Il n'existe pas d'estimateur adaptatif optimal pour ce problème. En d'autres termes, pour tout estimateur $\tilde{f}_\varepsilon$, il existe $(\beta_0, L_0) \in \mathcal{B}(\gamma)\times\mathcal{I}$ tel que :*

$$\limsup_{\varepsilon\to 0} \sup_{f\in H(\beta_0,L_0)} \mathbf{E}_f\left[\left(N_\varepsilon^{-1}(\beta_0, L_0)|\tilde{f}_\varepsilon(t) - f(t)|\right)^q\right] = +\infty.$$

Avant de donner le résultat principal concernant ce problème obtenu au cours de cette thèse, introduisons une notation.

$$\mathcal{H}(\gamma) = \bigcup_{(\beta,L)\in\mathcal{B}(\gamma)\times\mathcal{I}} H(\beta, L) = \bigcup_{\beta\in\mathcal{B}(\gamma)} H(\beta, l^*).$$

La connaissance de γ revient à faire l'hypothèse *a priori* que le signal inconnu appartient à $\mathcal{H}(\gamma)$.

Théorème 4. *Définissons la normalisation suivante :*

$$\eta_\varepsilon(\gamma) = (l^*)^{\frac{1}{2\gamma+1}} \left(\varepsilon\sqrt{\ln\ln\frac{1}{\varepsilon}}\right)^{\frac{2\gamma}{2\gamma+1}}.$$

C'est la vitesse minimax asympotique sur sur l'espace $\mathcal{H}(\gamma)$. On construit explicitement une procédure d'estimation qui atteint cette vitesse.

Les idées de la construction de cette procédure seront à la section suivante. Nous constatons par ailleurs que ce résultat prouve que la procédure adaptative, que nous avons construite, est optimale en un sens minimax. En effet, comme toutes les vitesses minimax, sur les différents espaces constituant la réunion $\mathcal{H}(\gamma)$ sont égales, l'approche minimax semble alors beaucoup plus raisonnable que dans le cas de l'adaptation totale où les vitesses étaient très différentes. C'est pourquoi nous avons adopté ce point de vue. De plus, le fait d'interprété notre résultat comme un résultat minimax sur une classe fonctionnelle inhabituelle nous a permis d'obtenir une vitesse de convergence inédite sur des classes de fonctions höldériennes.

Notons que dans ce cas d'adaptation avec information supplémentaire, la vitesse est meilleure que sans information. En effet la perte n'est que $\sqrt{\ln \ln 1/\varepsilon}$ alors qu'elle était de l'odre de $\sqrt{\ln 1/\varepsilon}$ dans le cas non informé.

Dans ce cas-ci, le prix à payer pour l'adaptation ne peut plus être compris comme un rapport de vitesses. On peut en revanche l'expliquer par le nombre élevé d'estimateurs qui entrent en jeu dans les comparaison pour construire la procédure. Plus précisément, on paie $\sqrt{\ln \#\text{nombre d'estimateurs}}$.

1.7 Perspectives

Dans cette thèse nous nous sommes intéressé à des problèmes d'estimation adaptative dans le modèle de bruit blanc gaussien où nous avons obtenu des résultats. Remarquons d'ailleurs que les deux problèmes que nous avons traités sont intimements liés l'un à l'autre. En effet, la borne inférieure concernantle résultat minimax prouve, en même temps, qu'il ne peut pas exister d'estimateur adaptatif optimal dans le problème d'estimation totale. Et, inversement, l'estimateur adaptatif total sert à construire l'estimateur minimax du deuxième problème.

Une des premières pistes de recherche est de prouver des résultats semblables dans les modèles moins idéaux que sont par exemple le modèle de densité ou le modèle de régression à pas aléatoire. On peut également s'intéresser au problème des constatntes exactes dans le modèle de bruit blanc gaussien.

Une autre piste de réflexion, plus théorique, est la suivante. En considérant un problème d'adaptation avec information supplémentaire, nous avons obtenu une vitesse minimax originale sur un espace fonctionnel. Peut-on, en

modifiant cette information disponible *a priori* obtenir d'autres vitesses originales ? Par exemple, si l'on arrivait à construire un sous-ensemble de $\mathcal{B}(\gamma)$ assez petit pour pouvoir construire une grille qui ne contienne que $\ln\ln 1/\varepsilon$ points, trouverait-on une vitesse minimax du type :

$$\left(\varepsilon\sqrt{\ln\ln\ln\frac{1}{\varepsilon}}\right)^{\frac{2\gamma}{2\gamma+1}} ?$$

Plus généralement, peut-on trouver des vitesses minimax (ou adaptatives ensuite) qui dépendent non seulement de la régularité des fonctions à estimer, mais également de la *géométrie* de l'hypothèse *a priori* ?

Enfin, on peut réinterpréter le résultat obtenu dans le cas de l'adaptation partielle de la manière suivante, un peu intermédiaire entre la théorie minimax et la théorie maxiset.

Pour commencer, on se donne un espace ambiant très gros. Dans notre cas il s'agit de $\mathcal{H} = \bigcup_{(\beta,L)\in\mathcal{B}\times\mathcal{I}} H(\beta, L)$. Ensuite, on fixe un précision qu'on souhaite atteindre, disons $\eta_\varepsilon(\gamma)$ avec $0 < \gamma < \bar{b}$. Le résultat que nous avons prouvé peut alors s'interpréter comme suit : nous avons construit simultanément un espace $\mathcal{H}(\gamma)$ et un estimateur $f_\varepsilon(\gamma)$ tels que :
– $f_\varepsilon(\gamma)$ atteint la précision $\eta_\varepsilon(\gamma)$ sur $\mathcal{H}(\gamma)$;
– $f_\varepsilon(\gamma)$ est minimax sur $\mathcal{H}(\gamma)$;
Pour avoir un résultat complet il faudrait encore montrer par exemple que $\mathcal{H}(\gamma)$ est un ensemble maximal (au sens de l'inclusion) vérifiant ces propriétés. C'est-à-dire que si $\mathcal{G}$ est tel que $\mathcal{H}(\gamma) \subsetneq \mathcal{G} \subset \mathcal{H}$, alors pour tout estimateur $\tilde{f}_\varepsilon(\cdot)$ on a :

$$\liminf_{\varepsilon\to 0} \sup_{f\in\mathcal{G}} \mathbf{E}_f\left[\left(\eta_\varepsilon^{-1}(\gamma)|\tilde{f}_\varepsilon(t) - f(t)|\right)^q\right] = +\infty.$$

Deuxième partie

Résultats

Chapter 2

Adaptive Procedure — Fully Case

2.1 Introduction

2.1.1 Model

In this paper, it is supposed that we observe the "trajectory" $\mathcal{X}^{(\varepsilon)} = \{X_\varepsilon(u)\}_{u\in\mathcal{D}}$ of a noisy signal which satisfies, on $[0,1]^d$ the following SDE:

$$X_\varepsilon(du) = \mathbf{f}(u)du + \varepsilon W(du)$$

where $\varepsilon > 0$ is a small parameter which represents the noise level, W is a Gaussian white noise on $[0,1]^d$ and $\mathbf{f}$ is the unknown signal to be estimated. It is assumed that $\mathbf{f}$ belongs to a given functional space $\Sigma(\varkappa)$ defined by a parameter $\varkappa$ which belongs to $\mathcal{J} \subset \mathbf{R}^m$.

Let us note that all results obtained in the paper remains valid if one replaces $[0,1]^d$ by an open set in $\mathbf{R}^d$.

Further, we will consider a particular case of this general framework where the functional space $\Sigma(\varkappa)$ is an anisotropic Hölder space $H(\beta, L)$. The exact definition of this space will be done later. Here, we mention only that $\beta = (\beta_1, \ldots, \beta_d)$ is an anisotropic (different in different directions) smoothness i.e. $\beta_i > 0$ represents the smoothness of the signal in the i^{th} direction and $L > 0$ is a Lipschitz constant. In this case $\varkappa = (\beta, L)$ and $m = d + 1$.

2.1.2 Quality of estimation. Minimax approach

Our goal is to estimate $\mathbf{f}$ at a given point $t \in (0,1)^d$. First, let us suppose that the "nuisance" parameter $\varkappa$ is known. To measure the quality of an arbitrary estimator $\tilde{f}_\varepsilon(\cdot) = \tilde{f}(\cdot\,; \mathcal{X}^{(\varepsilon)})$, we introduce its maximal risk on $\Sigma(\varkappa)$ as follows:

$$\forall q > 0, \quad R_\varepsilon^{(q)}\left[\tilde{f}_\varepsilon, \Sigma(\varkappa)\right] = \sup_{f \in \Sigma(\varkappa)} \mathbf{E}_f\left[\left|\tilde{f}_\varepsilon(t) - f(t)\right|^q\right].$$

We are interested in finding the asymptotic of the minimax risk (minimax rate of convergence):

$$N_\varepsilon^q(\varkappa) \asymp \inf_{\tilde{f}_\varepsilon} R_\varepsilon^{(q)}\left[\tilde{f}_\varepsilon, \Sigma(\varkappa)\right],$$

where the infimum is taken over all possible estimators.

Besides the finding $N_\varepsilon(\varkappa)$, we seek an estimator $\hat{f}_\varepsilon(\cdot)$ which achieves this rate i.e.

$$R_\varepsilon^{(q)}\left[\hat{f}_\varepsilon, \Sigma(\varkappa)\right] \asymp N_\varepsilon^q(\varkappa) \tag{U.B}$$

Any such estimator is called minimax.

Let us return to the anisotropic Hölder spaces. The solution of minimax problem was found in []. The minimax rate $N_\varepsilon(\beta, L)$ is given by the formula:

$$N_\varepsilon(\beta, L) = L^{1/(2\bar{\beta}+1)} \varepsilon^{2\bar{\beta}/(2\bar{\beta}+1)} \text{ where } 1/\bar{\beta} = \sum_{i=1}^{d} 1/\beta_i.$$

This rate is achieved by a kernel estimator with properly chosen kernel K and bandwidth $\eta = (\eta_1, \ldots, \eta_d)$. This estimator depends explicitly on (β, L) at least through its bandwidth defined by

$$\eta_i = \left(\frac{\varepsilon \|K\|}{L}\right)^{\frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i}}.$$

This is the typical situation: the solution of the minimax problem (rate of convergence, estimator) usually depends on the space where the minimax risk is defined.

2.1.3 Adaptive point of view

In practice, this dependence can be awkward. For instance, it is difficult to imagine that the smoothness β is exactly known. Usually, only the information on the belonging of the nuisance parameter to some set is available.

Formally, it is supposed that $\varkappa \in \mathcal{I} \subseteq \mathcal{J}$ or, in other words, $\mathbf{f}$ belongs to a known union of functional spaces.

Of course, we can adopt the minimax strategy: $\Sigma(\mathcal{I}) = \bigcup_{\varkappa\in\mathcal{I}} \Sigma(\varkappa)$ can be viewed as a new functional space. It is clear that the minimax on $\Sigma(\mathcal{I})$ estimator is independent on $\varkappa$. Let us note, nevertheless, that the minimax rate on this space, in other words the accuracy of minimax estimator, is not better than the "worse rate" $\sup_{\varkappa\in\mathcal{I}} N_\varepsilon(\varkappa)$. Therefore, if $\{N_\varepsilon(\varkappa)\}_{\varkappa\in\mathcal{I}}$ does not depend on $\varkappa$, this approach seems to be reasonable. For example, let us consider the family of anisotropic Hölder spaces $H(\beta, L), (\beta, L)$ such that $\bar{\beta} = \gamma$, then we have $N_\varepsilon(\beta, L) \asymp \varepsilon^{2\gamma/(2\gamma+1)}$ which is independent on β. On the other hand, in general situation, it is possible that for some $\varkappa$, $N_\varepsilon(\varkappa) \nrightarrow 0, \varepsilon \to 0$ or tends to 0 very slowly (in the previous example, it corresponds to the small values of the anisotropic smoothness β). Therefore, if $N_\varepsilon(\varkappa)$ is different (in order) for different values of $\varkappa$, this approach is not satisfactory.

Thus, we still seek a single estimator, but its accuracy should depend on the nuisance parameter $\varkappa$. Evidently we would like to have an estimator "as precise as possible" for each value of $\varkappa$. It leads to the first question arising in adaptive estimation. Does whether exist a single estimator that attains the minimax rate of convergence $N_\varepsilon(\varkappa)$ simultaneously on each space $\Sigma(\varkappa)$? Such estimator, if exists, is called optimal adaptive $\coprod$. Note that the accuracy given by this estimator cannot be improved for each value of $\varkappa$.

Unfortunately, optimal adaptive estimators do not always exist (Law, Tsybakov, Lepski). In this case we need a criterion of optimality in order to determine "the best estimator". To do it, we will follow the adaptive approach which consists in the following.

1. For any estimator $\tilde{f}_\varepsilon(\cdot)$, we consider the *family* of the normalized risks indexed by $\varkappa$

$$R_\varepsilon^{(q)}\left[\tilde{f}_\varepsilon, \Sigma(\varkappa), \psi_\varepsilon(\varkappa)\right] = \sup_{f\in\Sigma(\varkappa)} \mathbf{E}_f\left[\left(\psi_\varepsilon^{-1}(\varkappa)\left|\tilde{f}_\varepsilon(t) - f(t)\right|\right)^q\right], \ \varkappa \in \mathcal{I},$$

and let the family of normalizations $\Psi = (\psi_\varepsilon(\varkappa))_{\varkappa\in\mathcal{I}}$ and the estimator $f_\varepsilon^\Psi(\cdot)$ be such that

$$\limsup_{\varepsilon\to 0}\sup_{\varkappa\in\mathcal{I}} R_\varepsilon^{(q)}\left[\tilde{f}_\varepsilon^\Psi, \Sigma(\varkappa), \psi_\varepsilon(\varkappa)\right] < +\infty. \tag{A.U.B}$$

The family Ψ is called admissible.

2. We propose the criterion allowing to define the best admissible family of normalizations $\Phi = (\varphi_\varepsilon(\varkappa))_{\varkappa\in\mathcal{I}}$.

3. We construct an estimator f_ε^Φ satisfying (A.U.B) with $\Psi = \Phi$. This estimator will be called adaptive estimator.

Remark 1. *The main difficulty in realization of this program consists in finding a suitable criterion of optimality.*

- *The first attempt to give a satisfactory definition was undertaken in [].* *Then, it has been refined in [].* *In spite of the fact that these criteria can be applied to any statistical model, both of them are too rough in order to treat multidimensional problems.* *Below we present the criterion of optimality which generalizes the previous ones.*

- *The existence of an optimal adaptive estimator means that (A.U.B) is satisfied with* $\Psi = N \triangleq (N_\varepsilon(\varkappa))_{\varkappa\in\mathcal{I}}$, *i.e.* N *is admissible.* *In this case, any optimality criterion should guarantee that this family is optimal because it is impossible to improve* $N_\varepsilon(\varkappa)$ *for all* $\varkappa$.

2.1.4 Our results

In the present paper we study two different problems of adaptive estimation with respect to the collection of anisotropic Hölder spaces.

First, we consider the case when the nuisance parameter $\varkappa = (\beta, L)$ is completely unknown. In this case we find the optimal family of normalizations (in view of new criterion of optimality) and construct the adaptive estimator associated with this family (satisfying (A.U.B)). In particular, our result implies

- the optimal adaptive estimator does not exist for this estimation problem;

- the optimal family of normalization differs from the family of minimax rates of convergence $N_\varepsilon(\beta, L), (\beta, L) \in \mathcal{I}$, by a $\sqrt{\ln(1/\varepsilon)}$-factor that can be viewed as the price to pay for adaptation.

In dimension 1 the similar result was obtained in [] using another criterion of optimality. We replace it by finer criterion which is more suitable for multidimensional case.

It is worth to mention that our adaptive procedure is quite different from the estimator proposed in []. As in [], our estimator is a measurable choice from the collection of the kernel estimators but the strategy of the choice is much more sophisticated due to the dimension. Similar strategy was used in [].

Let us also note that proposed method is absolutely *parameter free*, and it is applied in the situation which we treat as "fully adaptive case".

The results discussed above form the first part of this paper.

Next, we suppose that the following additional information is available. The nuisance parameter $\varkappa = (\beta, L)$ is such that $1/\bar{\beta} = 1/\gamma$ where γ is given number. Let us make several remarks:

1. In this case the minimax rate of convergence on $H(\beta, L)$ does not depend on β and given by $\varepsilon^{2\gamma/(2\gamma+1)}$. This additional information can be treated as follows. We fix the desirable accuracy of estimation (choosing parameter γ) and look for an estimator providing it for any values of nuisance parameter (β, L). The important remark, here, is that the estimator attaining the rate $\varepsilon^{2\gamma/(2\gamma+1)}$ on $H(\beta, L)$ does not achieve it on $H(\alpha, L)$ for all $\alpha \neq \beta$, $1/\bar{\alpha} = 1/\gamma$. As, minimax rates do not depend on values of β, we can adopt the minimax strategy on the union of the anisotropic Hölder spaces $H(\beta, L)$ such that $1/\bar{\beta} = 1/\gamma$.

   We show that the minimax rate is asymptotically equivalent to

   $$\left(\varepsilon\sqrt{\ln\ln(1/\varepsilon)}\right)^{2\gamma/(2\gamma+1)}$$

   and construct the minimax estimator.

2. The construction of the minimax estimator (for given γ) uses the adaptive estimator obtained in the first part of the paper. This estimator could be called "partially adaptive" because the nuisance parameter (β, L) is unknown but not completely.

3. Note that found asymptotics differs from the minimax rate of convergence on $H(\beta, L), 1/\beta = 1/\gamma$ by the $\sqrt{\ln\ln(1/\varepsilon)}$-factor. It implies immediately that optimal adaptive estimators do not exist.

4. Finally, let us note that the additional information allows to minimize the price to pay for adaptation. As we mentioned before, this payment is $\sqrt{\ln(1/\varepsilon)}$ in the "fully adaptive case" and $\sqrt{\ln\ln(1/\varepsilon)}$ in the "partially adaptive case".

The partially adaptive problem forms the second part of this paper.

2.2 Basic definitions

2.2.1 Definition of the optimality

Motivations

As we already mentioned the first problem appearing in adaptive estimation is the existence of an optimal adaptive estimator (OAE). We will show that OAE with respect to the family of anisotropic Hölder spaces $\{H(\beta, L)\}_{(\beta,L)}$ does not exist. More precisely, we show that, for any estimator $\tilde{f}_\varepsilon$ there exists a value of the nuisance parameter, $(\beta_0, L_0) \in \mathfrak{J}$:

$$\limsup_{\varepsilon\to 0} \sup_{f\in H(\beta_0,L_0)} \mathbf{E}_f\left[\left(\varepsilon^{-\frac{2\bar{\beta}_0}{2\bar{\beta}_0+1}}\left|\tilde{f}_\varepsilon(t) - f(t)\right|\right)^q\right] = +\infty$$

Formally, this result means that the family of rates of convergence is not admissible.

In general case the non existence of an OAE can be formulated as follows: for any admissible family Ψ, there exists a nuisance parameter $\varkappa_0$ such that

$$\frac{\psi_\varepsilon(\varkappa_0)}{N_\varepsilon(\varkappa_0)} \underset{\varepsilon\to 0}{\longrightarrow} +\infty. \tag{2.1}$$

Clearly, it implies that any admissible family Ψ can be "improved", at least, in this point. In particular, one can use a minimax on $\Sigma(\varkappa_0)$ estimator for all values of nuisance parameter $\varkappa$. This estimator, which could be very bad

for all $\varkappa \neq \varkappa_0$, would outperform any estimator satisfying (AUB) with Ψ verifying (2.1).

As we see, the set of points where an admissible family can be improved is non empty. In this in mind, we will use the following principle in order to give the notion of optimality:

the "best admissible family" of normalizations should have "small number of points" where it can be improved.

Definition

Now, let us consider a general statistical experience $(V^\varepsilon, \mathcal{A}^\varepsilon, \{\mathbf{P}_f^\varepsilon\}_{f\in\Sigma})$ generated by the observation $\mathcal{X}^{(\varepsilon)}$, and let us suppose that $\Sigma = \bigcup_{\varkappa\in\mathcal{J}} \Sigma(\varkappa)$ where $\mathcal{J} \subset \mathbf{R}^m$ $(m \geq 1)$.

The goal is to estimate a functional $G(f)$ where $G : \Sigma \to (\Lambda, \|\cdot\|)$ where $(\Lambda, \|\cdot\|)$ is a Banach space.

In our particular case, let us recall that $(\Lambda, \|\cdot\|) = (\mathbf{R}, |\cdot|)$ and $G(f) = f(t)$. The maximal risk of an estimator $\tilde{f}_\varepsilon(\cdot)$ over the class $\Sigma(\varkappa)$ normalized by $\psi_\varepsilon(\varkappa)$ is defined by the formula

$$R_\varepsilon^{(q)}\left(\tilde{f}_\varepsilon, \Sigma(\varkappa), \psi_\varepsilon(\varkappa)\right) = \sup_{f\in\Sigma(\varkappa)} \mathbf{E}_f\left[\left(\psi_\varepsilon^{-1}(\varkappa)\|G(\tilde{f}_\varepsilon) - G(f)\|\right)^q\right].$$

This risk can be defined with a general loss function w satisfying usual assumptions and such that $w(u) \to +\infty, u \to +\infty$.

Let us introduce some definitions.

A family of normalizations $\Psi = (\psi_\varepsilon(\varkappa))_{\varkappa\in\mathcal{I}}$ is called admissible if there exists an estimator f_ε^Ψ such that the following inequality holds:

$$\limsup_{\varepsilon\to 0} \sup_{\varkappa\in\mathcal{I}} R_\varepsilon^{(q)}\left(f_\varepsilon^\psi, \Sigma(\varkappa), \psi_\varepsilon(\varkappa)\right) < +\infty.$$

For two admissible families Ψ and $\tilde{\Psi}$, we introduce two sets

$$\begin{aligned}
\mathcal{I}_0(\Psi/\tilde{\Psi}) &= \left\{\kappa \in \mathcal{I} : \quad \frac{\psi_\varepsilon(\kappa)}{\tilde{\psi}_\varepsilon(\kappa)} \xrightarrow[\varepsilon\to 0]{} 0\right\};\\
\mathcal{I}_\infty(\Psi/\tilde{\Psi}) &= \left\{\varkappa \in \mathcal{I} : \quad \frac{\psi_\varepsilon(\kappa)}{\tilde{\psi}_\varepsilon(\kappa)} \times \frac{\psi_\varepsilon(\varkappa)}{\tilde{\psi}_\varepsilon(\varkappa)} \xrightarrow[\varepsilon\to 0]{} +\infty,\ \forall \kappa \in \mathcal{I}_0(\Psi/\tilde{\Psi})\right\}.
\end{aligned}$$

The set $\mathcal{I}_0(\Psi/\tilde{\Psi})$ consists of all points where Ψ is "better", in order, than $\tilde{\Psi}$. One can say that $\tilde{\Psi}$ is "dominated" by Ψ on $\mathcal{I}_0(\Psi/\tilde{\Psi})$.

On the contrary, the set $\mathcal{I}_\infty(\Psi/\tilde{\Psi})$ consists of the points where $\tilde{\Psi}$ "dominates" Ψ and, moreover, the loss of $\tilde{\Psi}$ w.r.t. Ψ on $\mathcal{I}_0(\Psi/\tilde{\Psi})$ is "compensated".

Our principle of the choice between two admissible families Ψ and $\tilde{\Psi}$ consists in comparing of "massivities" of $\mathcal{I}_0(\Psi/\tilde{\Psi})$ and $\mathcal{I}_\infty(\Psi/\tilde{\Psi})$:

$\tilde{\Psi}$ *is "better" than* Ψ *if* $\mathcal{I}_\infty(\Psi/\tilde{\Psi})$ *is much more "massive" than* $\mathcal{I}_0(\Psi/\tilde{\Psi})$.

This idea leads to the following definition of an optimal family of normalizations.

Not to give the additional definitions, here and later, we will suppose that $\mathcal{I}$ contains an open set of $\mathbf{R}^m$.

Definition 1. *I) A family of normalizations* $\Phi = (\varphi_\varepsilon(\varkappa))_{\varkappa\in\mathcal{I}}$ *is called "optimal" if:*

i) Φ *is an "admissible" family.*

ii) If $\Psi = (\psi_\varepsilon(\varkappa))_{\varkappa\in\mathcal{I}}$ *is another admissible family of normalizations we have:*

- $\mathcal{I}_0(\Psi/\Phi)$ *is contained in a* $(m-1)$*-manyfold,*
- $\mathcal{I}_\infty(\Psi/\Phi)$ *contains an open set of* $\mathbf{R}^m$.

II) The estimator $f_\varepsilon^\Phi(\cdot)$ *is called an adaptive estimator.*

Let us comment this criterion.

Remark 2. *1. This definition is correct in the following sense:*

- *if* Φ *and* $\tilde{\Phi}$ *are two optimal families, then:*

$$\varphi_\varepsilon(\varkappa) \asymp \tilde{\varphi}_\varepsilon(\varkappa),\ \forall \varkappa \in \mathcal{I}.$$

- *If* N *is an admissible family (i.e. there exists an OAE), then it satisfies Definition 1. Indeed, in this case,* $\mathcal{I}_0(\Psi/N)$ *is empty, for any* Ψ.

2. *Note that we well followed our principle: the set of points where the optimal family Φ can be improved is really "small". Indeed the "dimension" of the set $\mathcal{I}_0(\Psi/\Phi)$ is* ***strictly*** *less than the dimension of $\mathcal{I}$ for* ***any*** *Ψ.*

3. *Let us also note that, the non existence of OAE implies that there exists Ψ such that $\mathcal{I}_0(\Psi/\Phi) \neq \emptyset$ i.e. there exists normalization (may be not unique) which "dominates" Φ on $\mathcal{I}_0(\cdot/\Phi)$. Let us denote $\mathfrak{N}$ the set of all normaliszations dominating Φ.*

   The message we would like to address is that the estimator f_ε^Φ (satisfying (AUB) with Φ) "outperforms" any estimator f_ε^Ψ satisfying (AUB) with Ψ belonging to $\mathfrak{N}$.

   Indeed, the estimator f_ε^Ψ is more precise than f_ε^Φ on $\mathcal{I}_0(\Psi/\Phi)$, which, let us remind is "very small set" for any $\Psi \in \mathfrak{N}$. The loss of f_ε^Φ w.r.t. f_ε^Ψ is given by

$$\{\varphi_\varepsilon(\varkappa)/\psi_\varepsilon(\varkappa) : \varkappa \in \mathcal{I}_0(\Psi/\Phi)\}.$$

   On the other hand, the estimator f_ε^Φ is more precise than f_ε^Ψ at least on $\mathcal{I}_\infty(\Psi/\Phi)$, which, in view of Definition 1, is very large. The gain of f_ε^Φ w.r.t. f_ε^Ψ is given, at least, by

$$\{\varphi_\varepsilon(\varkappa)/\psi_\varepsilon(\varkappa) : \varkappa \in \mathcal{I}_\infty(\Psi/\Phi)\}.$$

   In view of the definition of $\mathcal{I}_\infty(\cdot/\Phi)$, we can conclude that the gain of f_ε^Φ w.r.t. f_ε^Ψ (for ***any*** *$\Psi \in \mathfrak{N}$!), is much* bigger *on the* larger *set than its* loss *on the* smaller *set.*

2.2.2 Anisotropic Hölder spaces

To define the class of Hölder spaces let us introduce some notations. A function f belongs to $\mathcal{C}_\mathcal{D}$ if f is from $\mathbf{R}^d$ to $\mathbf{R}$ and it is compactly supported on $\mathcal{D}$. For a such function f, $i \in [\![1; d]\!]$ and $x \in \mathbf{R}^d$ we define:

$$\begin{aligned} f_i(\cdot|x) : \mathbf{R} &\to \mathbf{R} \\ y &\mapsto f(x_1, \dots, x_{i-1}, x_i + y, x_{i+1}, \dots, x_d) \end{aligned}$$

Let us denote $m_i(\beta) = \sup\{n \in \mathbf{N}; n < \beta_i\}$ and $\alpha_i(\beta) = \beta_i - m_i(\beta)$.

Definition 2. *Set $(\beta, L) \in \mathfrak{J}$. A function $f \in \mathcal{C}_\mathcal{D}$ belongs to the anisotropic Hölder space $H(\beta, L)$ if:*

- *The following property holds:*

$$\sup_{i=1,\dots,n} \sup_{x\in\mathbf{R}^d} \sum_{s=0}^{m_i} \left\| f_i^{(s)}(\cdot|x) \right\|_\infty \leq L,$$

- *for all $y \in \mathbf{R}$ and all $i \in [\![1; d]\!]$, the following inequality holds:*

$$\sup_{x\in\mathbf{R}^d} \left| f_i^{(m_i)}(y|x) - f_i^{(m_i)}(0|x) \right| \leq L|y|^{\alpha_i},$$

where $m_i = m_i(\beta)$ and $\alpha_i = \alpha_i(\beta)$.

In words, on the i^{th} direction of the canonical base of $\mathbf{R}^d$ the Hölder regularity (in the classical sense) of f is (β_i, L).

2.3 Our goal

Here and later, we consider the "fully adaptive problem".

Set $b = (b_1, \dots, b_d) \in (\mathbf{R}_+^*)^d$ and $l_* > 0$. Let us define, for all (β, L) such that $\bar{\beta} \leq \bar{b}$ and $L \geq l_*$, the following quantities:

$$\rho_\varepsilon(\beta, L) = \sqrt{1 + \frac{4(\bar{b} - \bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)} \ln \frac{L}{\|K\|\varepsilon} + \frac{2}{2\bar{b}+1} \ln \frac{L}{l_*}}$$

and

$$\varphi_\varepsilon(\beta, L) = L^{1/(2\bar{\beta}+1)} \left(\|K\|\varepsilon\rho_\varepsilon(\beta, L)\right)^{2\bar{\beta}/(2\bar{\beta}+1)}.$$

Here and later, it is assumed that $\varepsilon < l_*/\|K\|$. This assumption guarantees that $\rho_\varepsilon(\beta, L)$ (which will be viewed as the price to pay for adaptation) is greater than 1.

Let us denote $\Phi = (\varphi_\varepsilon(\beta, L))$. Our goal is to prove that Φ is the optimal family of normalizations w.r.t our criterion and, moreover, to construct an adaptive estimator, namely $f_\varepsilon^\Phi(\cdot)$, which satisfies (A.U.B) with Φ.

Remark 3. *At point (b, l_*), it is impossible to improve φ_ε since it corresponds at the minimax rate of convergence $N_\varepsilon(b, l^*)$. Let us also note that:*

$$\rho_\varepsilon(\beta, L) = \sqrt{1 + 2 \ln \frac{N_\varepsilon(\beta, L)}{N_\varepsilon(b, l_*)}}$$

2.4 Adaptive procedure

In this section, we describe the adaptive procedure. Let us recall that this procedure is constructed by the choice (data dependent) from the collection of† *kernel estimators*.

2.4.1 Kernels

A kernel is a function from $\mathbf{R}^d$ to $\mathbf{R}$ with some additional properties. We will denote $\mathcal{K}$ the class of kernel we will use. A kernel K belongs to $\mathcal{K}$ if it belongs to $\mathbf{L}^1(\mathbf{R}^d) \cap \mathbf{L}^2(\mathbf{R}^d)$ and satisfies the following properties:

$$\int_{\mathbf{R}^d} K(u)du = 1 \tag{K1}$$

$$\forall i \in [\![1; d]\!], \quad \int_{\mathbf{R}^d} |K(u)|(1 + |u_i|)^{b_i} du < +\infty \tag{K2}$$

$$\forall i \in [\![1; d]\!], \quad \int_{[-1,1]^d} |K(u)|du > 0. \tag{K3}$$

$$\forall i \in [\![1; d]\!], \forall l \in [\![1; b_i]\!], \quad \int_{\mathbf{R}^d} K(u)u_i^l du = 0. \tag{K4}$$

Further, "K is a kernel" will signify "K is a kernel belonging to $\mathcal{K}$". Then we will denote

$$\|K\| = \left(\int_{\mathbf{R}^d} |K(u)|^2 du \right)^{1/2}.$$

Remark 4. *Condition (K3) is a technical one. Other assumptions are classical.*

2.4.2 Collection of kernel estimators

First, for each $k \in \mathbf{Z}^d$ we define a bandwidth $h^{(k)} = (h_1^{(k)}, \dots, h_d^{(k)})$ in the following way. We introduce

$$h(b, l_*, \varepsilon) = \left(\frac{\|K\|\varepsilon}{l_*} \right)^{\frac{2\bar{b}}{2\bar{b}+1}\frac{1}{b_i}},$$

and, therefore

$$\forall i \in [\![1; d]\!], \quad h_i^{(k)} = h(b, l_*, \varepsilon) 2^{-(k_i+1)}.$$

Then, we can introduce, for all $k \in \mathbf{Z}^d$ the normalized kernel

$$K_k(u) = \left(\prod_{i=1}^{d} h_i^{(k)} \right)^{-1} K\left(\frac{u_1}{h_1^{(k)}}, \dots, \frac{u_d}{h_d^{(k)}} \right),$$

and the associated kernel estimator:

$$\hat{f}_k(t) = \int_{\mathbf{R}^d} K_k(t-u) X_\varepsilon(du).$$

Then, let us define

$$N_\varepsilon = \left\lfloor 2 \left(\frac{2\bar{b}}{2\bar{b}+1} \ln \frac{l_*}{\|K\|\varepsilon} + \ln \frac{l^*}{l_*} \right) \right\rfloor + 1$$

and

$$C(b) = \frac{2\bar{b}+1}{2\bar{b}} \times \frac{\ln 2 + \sqrt{2 \ln 2}}{\ln 2}.$$

For all $n \in [\![0; N_\varepsilon]\!]$, we consider the set

$$\mathcal{Z}(n) = \left\{ k =\in \mathbf{Z}^d : \sum_{i=1}^{d} (k_i + 1) = n \text{ and } \forall i, |k_i| \leq C(b) n + 1 \right\} \tag{2.2}$$

where $k = (k_1, \dots, k_d)$. This enables us to define

$$\mathcal{Z}_\varepsilon = \bigcup_{n=0}^{N_\varepsilon} \mathcal{Z}(n).$$

Finaly, we define the collection of estimators $\{\hat{f}_k(\cdot)\}_{k \in \mathcal{Z}_\varepsilon}$.

2.4.3 Procedure

Useful notations

Set

$$\begin{cases} \lambda_* &= \min\limits_{i \in [\![1;d]\!]} \int_{[-1,1]^d} |K(u)| \dfrac{|u_i|^{b_i}}{m_i(b)!} du \\ \lambda^* &= \max\limits_{i \in [\![1;d]\!]} \int_{\mathbf{R}^d} |K(u)| (1 + |u_i|)^{b_i} \end{cases}$$

and let

$$C = \frac{4}{\lambda_* d} + 2\sqrt{6q+4}.$$

Let us define a "partial ordering" on $\mathcal{Z}_\varepsilon$. We say that $k \preceq l$ if:

$$\sum_{i=1}^{d}(k_i+1) \triangleq |k| \leq |l| \triangleq \sum_{i=1}^{d}(l_i+1).$$

Let us also define, for all k and l in $\mathcal{Z}_\varepsilon$, $k \wedge l \in \mathbf{Z}^d$ by the formula:

$$k \wedge l = (k_1 \wedge l_1, \dots, k_d \wedge l_d).$$

The following quantities will be used throughout this paper.

$$\sigma_\varepsilon(l) = \frac{\varepsilon \|K\|}{\left(\prod_{i=1}^{d} h_i^{(l)}\right)^{1/2}}, \tag{2.3}$$

and

$$S_\varepsilon(l) = \sigma_\varepsilon(l)\sqrt{1 + |l| \ln 2}.$$

Remark 5. *Let us note that:*

$$\sigma_\varepsilon(l) = \sqrt{\mathrm{Var}\left(\hat{f}_l\right)}.$$

Adaptive estimator

Now, let us explain how the procedure chose an estimator. We introduce the random set $\mathcal{A}$ of all indexes defined by:

$$\mathcal{A} = \left\{ k \in \mathcal{Z}_\varepsilon : \left| \hat{f}_{k\wedge l}(t) - \hat{f}_l(t) \right| \leq C S_\varepsilon(l),\ \forall l \in \mathcal{Z}_\varepsilon,\ l \succeq k \right\}.$$

If $\mathcal{A}$ is non empty, one can chose $\hat{k}$ (may be not unique) such that:

$$\hat{k} = \arg\min_{k \in \mathcal{A}} \sigma_\varepsilon(k).$$

Remark that this choice of $\hat{k}$ is measurable because $\mathcal{A}$ is a finite set ($\mathcal{A} \subseteq \mathcal{Z}_\varepsilon$, $\mathcal{Z}_\varepsilon$ finite). Now, we can construct explicitly our estimator f_ε^Φ in the following way:

$$f_\varepsilon^\Phi(t) = \begin{cases} \hat{f}_{\hat{k}}(t) & \text{if } \mathcal{A} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

2.4.4 Comments

Let us make several comments about this procedure.

- Our procedure is quite different to that introduced in [] to solve the similar one-dimensional problem. The main difference is connected with the manner of choosing a random index $\hat{k}$. Here, we compare *pairwise* the estimators by introducing the "artificial" estimator $\hat{f}_{k\wedge l}(\cdot)$ in definition of $\mathcal{A}$. Then, we chose an estimator of minimal variance estimators $\{\hat{f}_k\}_{k\in\mathcal{A}}$.

  In dimension 1, it is useless to introduce these artificial estimators. Unfortunately, a such procedure is not adapted to solve multidimensional problems (if one deals with anisotropic regularities) and fails.

- Our procedure is inspired by the method proposed in [], well adapted to multidimensional problems. Let us mention however, that it is not possible to use this method directly. The main difference is the choice of set of indexes. In our case, we have to consider (it will be explained further in this paper) indexes belonging to $\mathbf{Z}^d$ — instead of $\mathbf{N}^d$ considered in []. First of all, let us remind that our set of indexes is

$$\mathcal{Z}_\varepsilon = \bigcup_{n=0}^{N_\varepsilon} \mathcal{Z}(n) \subseteq \mathbf{Z}^d.$$

  The set used in [] is $\mathcal{N}_\varepsilon = \bigcup_n \mathcal{N}(n)$ where:

$$\mathcal{N}(n) = \left\{ k = (k_1, \dots, k_d) \in \mathbf{N}^d : \sum_{i=1}^{d} (k_i + 1) = n \right\}.$$

  Both procedures require the following properties of indexes:

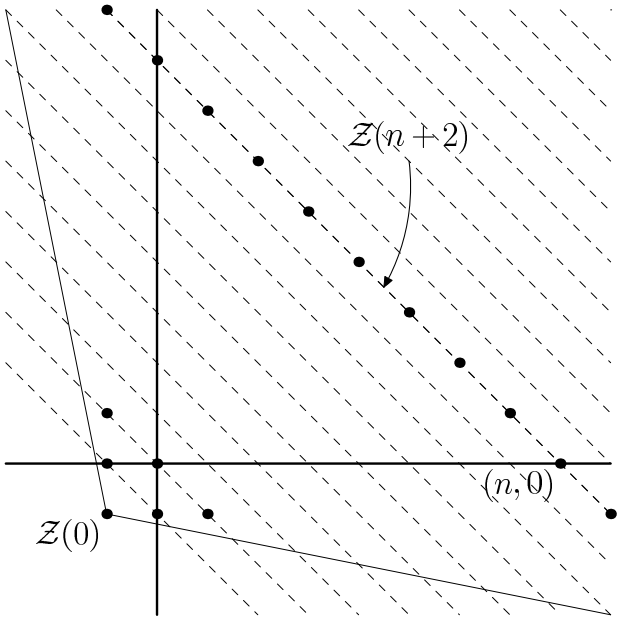
$$\begin{cases} \displaystyle\sum_{n=0}^{\infty} 2^{-n} \#\mathcal{Z}(n) & < \infty \\ \displaystyle\sum_{n=0}^{\infty} 2^{-n} \#\mathcal{N}(n) & < \infty. \end{cases} \tag{2.4}$$

  Second property is evidently fulfilled. The first one requires a special construction given by (2.2).

  Let us also note that not to pay an additional price at final point (b, l_*), we need $\mathcal{Z}(0)$ is bounded independently on ε which follows immediately from 2.4 otherwise we need to pay $\sqrt{1 + \ln \mathcal{Z}(0)}$.

Remark 6. *Figure 1 represents $\mathcal{Z}_\varepsilon$ in dimension 2. Here, $\mathcal{Z}(i), i = 0, 1, 2$ and $\mathcal{Z}(n+2)$ are drawn.The black points belong to $\mathcal{Z}_\varepsilon$.*

Figure 2.1: $\mathcal{Z}_\varepsilon$.



2.4.5 Upper bound

Let us introduce some basic notations: $b = (b_1, \dots, b_d)$ is a vector of positive numbers and $0 < l_* < l^* < +\infty$ are given. Set

$$\mathcal{B} = \prod_{i=1}^{d} (0, b_i] \text{ and } \mathcal{L} = [l_*, l^*].$$

Moreover, for all $\gamma \in (0, \bar{b}]$, let us consider

$$\mathcal{B}(\gamma) = \left\{ \beta \in \mathcal{B}; 1/\bar{\beta} = 1/\gamma \right\}.$$

Let us denote

$$\mathfrak{J} = \mathcal{B} \times \mathcal{L} \text{ and } \mathfrak{I}(\gamma) = \mathcal{B}(\gamma) \times \mathcal{L}.$$

Theorem 1. *Set $\varepsilon < l_*/\|K\|$ and $q > 0$. Then*

$$\sup_{(\beta,L)\in\mathfrak{J}} R_\varepsilon^{(q)}\left(f_\varepsilon^\Phi(t), H(\beta,L), \varphi_\varepsilon(\beta,L)\right) \leq M_q$$

where M_q is an absolute constant which does not depend on ε. The explicit expression of $M_q = M_q(l_*, l^*, b)$ is given in the proof.

Remark 7. *This result implies clearly that f_ε^Φ satisfies (A.U.B) with Φ but it is stronger: in fact, we obtain a non-asymptotical upper bound for all ε small enough.*

2.5 Optimality of Φ

2.5.1 Result

Theorem 2. *Set $\psi = (\psi_\varepsilon(\beta,L))_{(\beta,L)\in\mathfrak{J}}$ an admissible family of normalizations such that $\exists \beta_0 \in \mathcal{I}_0(\Psi/\Phi)$, then:*

1. *$\mathcal{I}_0(\Psi/\Phi) \subseteq \mathfrak{I}(\bar{\beta}_0)$;*
2. *$\mathcal{I}_\infty(\Psi/\Phi) \supseteq \bigcup_{\gamma>\bar{\beta}_0} \mathfrak{I}(\gamma)$.*

Remark 8. *This result implies that Φ is the optimal family w.r.t to our criterion. Indeed, $\dim(\mathfrak{I}(\bar{\beta}_0)) = d < d+1 = \dim(\mathfrak{J})$ and it is clear that $\bigcup_{\gamma>\bar{\beta}_0} \mathfrak{I}(\gamma)$ contains an open set of $\mathfrak{J}$.*

2.5.2 Comments

Let us briefly discuss an interesting point which "shows" that our criterion of optimality is well adapted to our problem. One of our idea, by introducing this criterion, was to minimize $\mathcal{I}_0(\Psi/\Phi)$ (in term of massivity). Theorem 2 says that this set is always contained in $\mathfrak{I}(\gamma)$ for a given γ. Can we improve this result (by proving that $\mathcal{I}_0(\Psi/\Phi)$ is essentially smaller than $\mathfrak{I}(\gamma)$)? The answer is no!

Indeed, let us suppose that the result concerning the "partially adaptive problem" is proved. Thus, for all $0 < \gamma < \overline{b}$, the minimax rate of convergence on

$$\mathcal{F}(\gamma) = \bigcup_{(\beta,L)\in\mathfrak{I}(\gamma)} H(\beta,L)$$

is given by

$$\phi_\varepsilon(\gamma) \asymp \left(\varepsilon\sqrt{\ln\ln\frac{1}{\varepsilon}}\right)^{\frac{2\gamma}{2\gamma+1}}.$$

It is evident that any estimator which achieves this rate on $\mathcal{F}(\gamma)$ outperform f_ε^Φ at least on $\mathfrak{I}(\gamma)$. The loss is about:

$$\left(\frac{\ln 1/\varepsilon}{\ln\ln 1/\varepsilon}\right)^{\frac{\gamma}{2\gamma+1}}$$

Combining this result with Theorem 2, we obtain

$$\mathcal{I}_0(\Psi^\gamma/\Phi) = \mathfrak{I}(\gamma).$$

where $\Psi^\gamma = (\psi_\varepsilon^\gamma(\beta, L))_{(\beta,L)}$ is defined by

$$\psi_\varepsilon^\gamma(\beta, L) = \begin{cases} \phi_\varepsilon(\gamma) & \text{if } \bar{\beta} = \gamma \\ 1 & \text{otherwise.} \end{cases}$$

2.6 Proof of theorem 1

2.6.1 Introduction

Let us explain, briefly, the main ideas to prove our result.

First, let us suppose that the smoothness parameter (β, L) of the signal (to be estimated) is well known. Thus, it is easy to construct an estimator (depending on (β, L)) witch achieves the expected rate $\varphi_\varepsilon(\beta, L)$.

To do that, we have to chose $\tilde{h}(\beta, L, \varepsilon) = (\tilde{h}_1(\beta, L, \varepsilon), \ldots, \tilde{h}_d(\beta, L, \varepsilon))$ (bandwidth of this estimator) on the following way:

$$\tilde{h}_i(\beta, L, \varepsilon) = \gamma_i(\beta)\left(\frac{\|K\|\Gamma(\beta)}{2L}\varepsilon\rho_\varepsilon(\beta, L)\right)^{\frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i}}, \forall i,$$

where

$$\begin{cases} \gamma_i(\beta) &= (\lambda_i(\beta)\beta_i)^{-1/\beta_i} \\ \Gamma(\beta) &= \left(\prod_{i=1}^d \gamma_i(\beta)\right)^{-1/2}. \end{cases}$$

This formula is obtained as the solution of the following minimization problem:

$$\tilde{h}(\beta, L, \varepsilon) = \arg\min_{h \in \mathcal{H}} \left(b^{\beta,L} + s_\varepsilon^{\beta,L}\right)(h) \tag{2.5}$$

where

$$b^{\beta,L}(h) = L \sum_{i=1}^{d} \lambda_i(\beta) h_i^{\beta_i}$$

is a *bias* term and

$$s_\varepsilon^{\beta,L}(h) = \frac{\|K\|\varepsilon}{\sqrt{\prod_{i=1}^{d} h_i}} \rho_\varepsilon(\beta, L)$$

can be viewed as a *penalized standard deviation* term.

Remark 9. *Using these notations we obtain*

$$b^{\beta,L}(\tilde{h}(\beta, L, \varepsilon)) \asymp s_\varepsilon^{\beta,L}(\tilde{h}(\beta, L, \varepsilon)) \asymp \varphi_\varepsilon(\beta, L), \quad \forall (\beta, L).$$

Next, if (β, L) is unknown, we want that our procedure choses a kernel estimator as good as the optimal one, constructed using bandwidth $\tilde{h}(\beta, L, \varepsilon)$. In order to do that, our procedure compare a large number of estimators. In particular, for each $(\beta, L) \in \mathfrak{J}$, the estimator constructed using bandwidth $\tilde{h}(\beta, L, \varepsilon)$ should be "viewed" by the procedure. This implies that set $\mathcal{Z}_\varepsilon$ is large enough.

2.6.2 Lemmas

Here, we give some lemmas witch will be proved in Appendix. They are used further in the proof.

Lemma 1. *Set* (β, L).

Bandwidth $\tilde{h}(\beta, L, \varepsilon)$ *is the unique bandwidth* $\tilde{\eta}$ *such that:*

$$\tilde{\eta} = \arg\min_{h \in \mathcal{H}} \left(b^{\beta,L} + s_\varepsilon^{\beta,L}\right)(h).$$

For simplicity, let us denote $h(\beta, L, \varepsilon) = (h_1(\beta, L, \varepsilon), \dots, h_d(\beta, L, \varepsilon))$ defined by:

$$h_i(\beta, L, \varepsilon) = \left(\frac{\|K\|}{L} \varepsilon \rho_\varepsilon(\beta, L)\right)^{\frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i}}.$$

It is clear that, the estimator defined using bandwidth $h(\beta, L, \varepsilon)$ is asymptoticaly as good as the estimator defined using bandwidth $\tilde{h}(\beta, L, \varepsilon)$. Indeed, for all (β, L) we have:

$$h_i(\beta, L, \varepsilon) \asymp \tilde{h}_i(\beta, L, \varepsilon), \quad \forall i.$$

Now, let us consider

$$k_i(\beta, L, \varepsilon) = \left\lfloor \frac{1}{\ln 2} \ln \left(\frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \right) \right\rfloor$$

where $\lfloor x \rfloor = \sup\{n \in \mathbf{N} : n \leq x\}$. And let us consider the index $k(\beta, L, \varepsilon) = (k_1(\beta, L, \varepsilon), \dots, k_d(\beta, L, \varepsilon))$ in $\mathbf{Z}^d$.

It is easy to see that the estimator defined by the bandwidth $h^{(k(\beta,L,\varepsilon))}$ is asymptoticaly as good as the estimator defined by $h(\beta, L, \varepsilon)$ and, thus, as good as that one defined by $\tilde{h}(\beta, L, \varepsilon)$.

Lemma 2. *Set (β, L). Index $k(\beta, L, \varepsilon)$ belongs to $\mathcal{Z}_\varepsilon$.*

Remark 10. *Set $\mathcal{Z}_\varepsilon$ was constructed such that this lemma is satisfied and moreover such that inequality (2.4) holds.*

Let us give an improtant lemma concerning the canonical decomposition of the estimator $\hat{f}_k$.

Lemma 3. *Let us fix $f \in \bigcup_{(\beta,L)\in\mathcal{B}\times\mathcal{I}} H(\beta, L)$, and let us calculate under the law $\mathbf{P}_f$. We have, for $k \in \mathcal{Z}_\varepsilon$:*

$$\hat{f}_k(t) = f(t) + b_k(t) + \sigma_\varepsilon(k)\xi(k),$$

where:

$$\begin{cases} b_k(t) &= \displaystyle\int_{\mathbf{R}^d} K(u) \left(f(t - h^{(k)}.u) - f(t) \right) du \\ \sigma_\varepsilon(k) &= \dfrac{\|K\|\varepsilon}{\left(\prod_{i=1}^d h_i^*(\varepsilon) \right)^{1/2}} 2^{\frac{|k|}{2}} \\ \xi(k) &\sim \mathcal{N}(0, 1), \end{cases}$$

where $h.u$ denotes the following vector: $(h_1 u_1, \dots, h_d u_d)$. Moreover, let us remark that, if $k \preceq l$, then $\sigma_\varepsilon(k) \leq \sigma_\varepsilon(l)$.

Now, let us give the most important lemma about the control of bias terms. More precisely:

Lemma 4. *Set $(\beta, L) \in \mathcal{B} \times \mathcal{I}$ and $f \in H(\beta, L)$. Under $\mathbf{P}_f$ we have:*

$$\forall k \in \mathcal{Z}_\varepsilon, \quad |b_k(t)| \leq B^{\beta,L}(k)$$

and

$$\forall (k,l) \in \mathcal{Z}_\varepsilon^2, \quad |b_{k\wedge l}(t) - b_l(t)| \leq 2B^{\beta,L}(k),$$

where $B^{\beta,L}(k) = b^{\beta,L}(h^{(k)})$.

Now, let us give a lemma concerning the link beetwen the bias and the penalized standard deviation of the estiamtor $\hat{f}_{k(\beta,L,\varepsilon)}$. First of all let us recall that $S_\varepsilon(k)$ was define by equation (2.3).

Lemma 5. *For all $(\beta, L) \in \mathcal{B} \times \mathcal{I}$ we have:*

$$B^{\beta,L}(k(\beta, L, \varepsilon)) \leq C^* S_\varepsilon(k(\beta, L, \varepsilon)),$$

where $C^* = (d\lambda^*)\sqrt{2 \vee (\bar{b}+1)/\underline{b}}$.

Finally, let us give a lemma which explain a link beetwen $S_\varepsilon(k(\beta, L, \varepsilon))$ and the rate of convergence $\varphi_\varepsilon(\beta, L)$. It is very important, because this lemma proves that it is enough to control (up to a constant) the quality of the estimator f_ε^Φ by $S_\varepsilon(k(\beta, L, \varepsilon))$.

Lemma 6. *For all $(\beta, L) \in \mathcal{B} \times \mathcal{I}$, we have:*

$$S_\varepsilon(k(\beta, L, \varepsilon)) \leq \varphi_\varepsilon(\beta, L).$$

Lemma 3 is evident. All the others will be proved in Apendix.

2.6.3 Proof

Let us consider $\varepsilon < l_*/\|K\|$ and $q > 0$.

We want to prove that

$$\sup_{(\beta,L)\in\mathfrak{J}} \sup_{f\in H(\beta,L)} \mathbf{E}_f\left[\left(\varphi_\varepsilon^{-1}(\beta, L)\left|f_\varepsilon^\Phi(t) - f(t)\right|\right)^q\right] < +\infty.$$

Thus, we fix $(\beta, L) \in \mathfrak{J}$ and $f \in H(\beta, L)$. Let us denote $\kappa = k(\beta, L, \varepsilon)$. Let us recall that $k(\beta, L, \varepsilon)$ is the index correponding to the bandwidth $h(\beta, L, \varepsilon)$.

Now, we have to distinguish two cases: First $\mathcal{A}$ is empty. Next, it is non empty.

A) $\mathcal{A}$ is non empty

In this case, the procedure chose an index $\hat{k}$. Main idea is the following: we have to compare $\hat{f}_{\hat{k}}$ and $\hat{f}_{\kappa}$. To do that, let us introduce the following sets, for all $s \in \mathbf{N}$:

$$\begin{cases} B_1(s) &= \{k \in \mathcal{Z}_\varepsilon : |k| \leq |\kappa| + sd\} \\ B_2(s) &= \{k \in \mathcal{Z}_\varepsilon : |k| > |\kappa| + (s-1)d\} \end{cases}$$

Using these notations we obtain:

$$\mathcal{Z}_\varepsilon = B_1(0) \cup \left(\bigcup_{s \geq 1} B_1(s) \cap B_2(s) \right).$$

Thus, we have:

$$\begin{aligned} \mathbf{E}_f\left[\left(\left|f_\varepsilon^\Phi(t) - f(t)\right|\right)^q\right] &\leq \mathbf{E}_f\left[\left|\hat{f}_{\hat{k}}(t) - f(t)\right|^q \mathbf{1}_{\{\hat{k} \in B_1(0)\}}\right] \\ &\quad + \sum_{s \geq 1} \mathbf{E}_f\left[\left|\hat{f}_{\hat{k}}(t) - f(t)\right|^q \mathbf{1}_{\{\hat{k} \in B_1(s)\}} \mathbf{1}_{\{\hat{k} \in B_2(s)\}}\right] \\ &\leq R(0,q) + \sum_{s \geq 1} \sqrt{R(s,2q)D(s)}, \end{aligned}$$

where

$$\begin{cases} R(s,p) = \mathbf{E}_f\left[\left|\hat{f}_{\hat{k}}(t) - f(t)\right|^p \mathbf{1}_{\{\hat{k} \in B_1(s)\}}\right] \\ D(s) = \mathbf{P}_f\left[\hat{k} \in B_2(s)\right]. \end{cases}$$

Thus we have to control these quantities.

Control of $D(s)$. Let us denote $\kappa(s) = (\kappa_1 + s, \ldots, \kappa_d + s)$ and let us consider $s_{\max} = \max\{s \in \mathbf{N} : B_2(s) \neq \emptyset\}$. Clearly $s_{\max} \leq N_\varepsilon + 1$. Thus we have $D(s) = 0$ for any $s > s_{\max}$. Let us consider $s \leq s_{\max}$. It is easy to see that $\hat{k} \in B_2(s) \Rightarrow \kappa(s-1) \notin \mathcal{A}$. Thus, if $\hat{k} \in B_2(s)$, then there exists $l \in \mathcal{Z}_\varepsilon, l \succeq \kappa(s-1)$, such that

$$\left|\hat{f}_{\kappa(s-1)\wedge l}(t) - \hat{f}_l(t)\right| > CS_\varepsilon(l).$$

Let us denote

$$D_l(s) = \mathbf{P}_f\left[\left|\hat{f}_{\kappa(s-1)\wedge l}(t) - \hat{f}_l(t)\right| > CS_\varepsilon(l)\right].$$

Using this notation it follows:

$$D(s) \leq \sum_{l \in \mathcal{Z}_\varepsilon, l \succeq \kappa(s-1)} D_l(s).$$

Now, we have to control $D_l(s)$. Set $l \succeq \kappa(s-1)$. We have, using lemmas (3), (4) and (5):

$$\begin{aligned} D_l(s) &\leq \mathbf{P}_f\big[|b_{\kappa(s-1)\wedge l}(t) - b_l(t)| \\ &\qquad +\sigma_\varepsilon(\kappa(s-1)\wedge l)|\xi(\kappa(s-1)\wedge l)| \\ &\qquad +\sigma_\varepsilon(l)|\xi(l)| > CS_\varepsilon(l)\big] \\ &\leq \mathbf{P}_f\big[2B^{\beta,L}(\kappa(s-1)) \\ &\qquad +\sigma_\varepsilon(\kappa(s-1)\wedge l)|\xi(\kappa(s-1)\wedge l)| \\ &\qquad +\sigma_\varepsilon(l)|\xi(l)| > CS_\varepsilon(l)\big] \\ &\leq \mathbf{P}_f\big[2C^*S_\varepsilon(\kappa(s-1)) \\ &\qquad +\sigma_\varepsilon(\kappa(s-1)\wedge l)|\xi(\kappa(s-1)\wedge l)| \\ &\qquad +\sigma_\varepsilon(l)|\xi(l)| > CS_\varepsilon(l)\big]. \end{aligned}$$

Using lemma (3), it follows

$$\begin{aligned} D_l(s) &\leq \mathbf{P}_f\Big[|\xi(\kappa(s-1)\wedge l)| + |\xi(l)| > (C-2C^*)\sqrt{1+|l|\ln 2}\Big] \\ &\leq 2\mathbf{P}\left[|\mathcal{N}(0,1)| > \frac{C-2C^*}{2}\sqrt{1+|l|\ln 2}\right] \\ &\leq 2^{-\tilde{C}|l|+1}, \end{aligned}$$

where $\tilde{C} = (C-2C^*)^2/8$. Thus,

$$D(s) \leq 2\sum_{l\in\mathcal{Z}_\varepsilon, l\succeq\kappa(s-1)} 2^{-\tilde{C}|l|}. \tag{2.6}$$

Control of $R(s,p)$. Let us recall that

$$R(s,p) = \mathbf{E}_f\left[\left|\hat{f}_{\hat{k}}(t) - f(t)\right|^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right].$$

The main idea is to decompose

$$\left|\hat{f}_{\hat{k}}(t) - f(t)\right|$$

by introducing $\hat{f}_{\kappa(s)}(t)$. In order to do that, we have to introduce $\hat{f}_{\hat{k}\wedge\kappa(s)}(t)$. Let us write:

$$\begin{aligned} \left|\hat{f}_{\hat{k}}(t) - f(t)\right| &\leq \left|\hat{f}_{\hat{k}}(t) - \hat{f}_{\hat{k}\wedge\kappa(s)}(t)\right| \\ &\quad + \left|\hat{f}_{\hat{k}\wedge\kappa(s)}(t) - \hat{f}_{\kappa(s)}(t)\right| \\ &\quad + \left|\hat{f}_{\kappa(s)}(t) - f(t)\right|. \end{aligned}$$

It is easy to prove that, if $s \leq s_{\max}$ then $\kappa(s)$ belongs to $\mathcal{Z}_\varepsilon$. This is a very important point. We will consider only $s \leq s_{\max}$. Let us recall that, if $s > s_{\max}$ then $D(s) = 0$.

Let us denote:

$$\begin{cases} I_1 &= |\hat{f}_{\hat{k}}(t) - \hat{f}_{\hat{k}\wedge\kappa(s)}(t)|\mathbf{1}_{\{\hat{k}\in B_1(s)\}} \\ I_2 &= |\hat{f}_{\hat{k}\wedge\kappa(s)}(t) - \hat{f}_{\kappa(s)}(t)|\mathbf{1}_{\{\hat{k}\in B_1(s)\}} \\ I_3 &= |\hat{f}_{\kappa(s)}(t) - f(t)|\mathbf{1}_{\{\hat{k}\in B_1(s)\}}. \end{cases}$$

We have:

$$R(s,p) \leq \left(3^{p-1} \vee 1\right)\left(\mathbf{E}_f[I_1^p] + \mathbf{E}_f[I_2^p] + \mathbf{E}_f[I_3^p]\right).$$

a) Let us control $\mathbf{E}_f[I_3^p]$. Using lemmas (3), (4) and (5), we have:

$$\begin{aligned} \mathbf{E}_f[I_3^p] &= \mathbf{E}_f\left[|\hat{f}_{\kappa(s)}(t) - f(t)|\mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \\ &\leq \mathbf{E}_f\left[\left(|b_{\kappa(s)}(t)| + \sigma_\varepsilon(\kappa(s))|\xi(\kappa(s))|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \\ &\leq \mathbf{E}_f\left[\left(B^{\beta,L}(\kappa(s)) + \sigma_\varepsilon(\kappa(s))|\xi(\kappa(s))|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \\ &\leq \mathbf{E}_f\left[\left(C^* S_\varepsilon(\kappa(s)) + \sigma_\varepsilon(\kappa(s))|\xi(\kappa(s))|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right]. \end{aligned}$$

Thus, we obtain:

$$\begin{aligned} \mathbf{E}_f[I_3^p] \leq\ & \left(2^{p-1} \vee 1\right)(C^*)^p S_\varepsilon^p(\kappa(s)) \\ & + \left(2^{p-1} \vee 1\right)\mathbf{E}_f\left[\left(\sigma_\varepsilon(\kappa(s))|\xi(\kappa(s))|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right]. \end{aligned} \tag{2.7}$$

b) Let us control $\mathbf{E}_f[I_2^p]$. First, let us remark that:

- $\hat{k}$ belongs to $\mathcal{A}$. By definition of $\hat{k}$.
- $|\kappa(s)| \geq |\hat{k}|$. Because $\hat{k}$ belongs to $B_1(s)$.
- $\kappa(s)$ belongs to $\mathcal{Z}_\varepsilon$. Thanks to lemma **??**.

Thus, the construction of our procedure implies that

$$\mathbf{E}_f[I_2^p] \leq C^p S_\varepsilon^p(\kappa(s)). \tag{2.8}$$

c) Let us control $\mathbf{E}_f[I_1^p]$. Using lemmas 3, 4 and 5, it is easy to see that:

$$\begin{aligned} I_1 \leq\ & 2C^* S_\varepsilon(\kappa(s))\mathbf{1}_{\{\hat{k}\in B_1(s)\}} \\ & +\sigma_\varepsilon(\hat{k})|\xi(\hat{k})|\mathbf{1}_{\{\hat{k}\in B_1(s)\}} \\ & +\sigma_\varepsilon(\hat{k}\wedge\kappa(s))|\xi(\hat{k}\wedge\kappa(s))|\mathbf{1}_{\{\hat{k}\in B_1(s)\}}. \end{aligned}$$

Thus, we obtain:

$$\begin{aligned} \mathbf{E}_f[I_1^p] \leq\ & \left(3^{p-1}\vee 1\right)(2C^*)^p S_\varepsilon^p(\kappa(s)) \\ & + \left(3^{p-1}\vee 1\right)\mathbf{E}_f\left[\left(\sigma_\varepsilon(\hat{k})|\xi(\hat{k})|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \\ & + \left(3^{p-1}\vee 1\right)\mathbf{E}_f\left[\left(\sigma_\varepsilon(\hat{k}\wedge\kappa(s))|\xi(\hat{k}\wedge\kappa(s))|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \quad (2.9) \end{aligned}$$

Using inequalities (2.7)–(2.8)–(2.9), we obtain:

$$\begin{aligned} R(s,p) \leq\ & \left(\left(2^{p-1}\vee 1\right)(C^*)^p + C^p + \left(3^{p-1}\vee 1\right)(2C^*)^p\right) S_\varepsilon^p(\kappa(s)) \\ & + \left(2^{p-1}\vee 1\right)\mathbf{E}_f\left[\left(\sigma_\varepsilon(\kappa(s))|\xi(\kappa(s))|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \\ & + \left(3^{p-1}\vee 1\right)\mathbf{E}_f\left[\left(\sigma_\varepsilon(\hat{k})|\xi(\hat{k})|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \\ & + \left(3^{p-1}\vee 1\right)\mathbf{E}_f\left[\left(\sigma_\varepsilon(\hat{k}\wedge\kappa(s))|\xi(\hat{k}\wedge\kappa(s))|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \quad (2.10) \end{aligned}$$

Thus, we have to control the expectations in the last inequality.

It is easy to control the first one:

$$\mathbf{E}_f\left[\left(\sigma_\varepsilon(\kappa(s))|\xi(\kappa(s))|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \leq \sigma_\varepsilon(\kappa(s))\mathbf{E}\left[|\mathcal{N}(0,1)|^p\right].$$

Now, let us explain how to control the others. Let us denote $\tilde{k} = \hat{k}\wedge\kappa(s)$ and

$$\Lambda_k = \left\{|\xi(k\wedge\kappa(s))| > 2\sqrt{1+|k|\ln 2}\right\}.$$

Now, let us calculate:

$$\begin{aligned} (*) =\ & \mathbf{E}_f\left[\left(\sigma_\varepsilon(\hat{k}\wedge\kappa(s))|\xi(\hat{k}\wedge\kappa(s))|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \\ =\ & \mathbf{E}_f\left[\left(\sigma_\varepsilon(\tilde{k})|\xi(\tilde{k})|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\left(\mathbf{1}_{\{\Lambda_{\hat{k}}\}} + \mathbf{1}_{\{\Lambda_{\hat{k}}^c\}}\right)\right] \\ \leq\ & \mathbf{E}_f\left[\left(\sigma_\varepsilon(\tilde{k})|\xi(\tilde{k})|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\mathbf{1}_{\{\Lambda_{\hat{k}}\}}\right] \\ & + \mathbf{E}_f\left[\left(2\sigma_\varepsilon(\tilde{k})\sqrt{1+|\hat{k}|\ln 2}\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \\ \leq\ & \sigma_\varepsilon^p(\kappa(s))\mathbf{E}_f\left[|\xi(\tilde{k})|^p \mathbf{1}_{\{\Lambda_{\hat{k}}\}}\right] + \left(2S_\varepsilon(\kappa(s))\right)^p. \end{aligned}$$

Let us denote $m_l = \mathbf{E}\left[|\mathcal{N}(0,1)|^l\right]$. We obtain:

$$(*) \leq \sigma_\varepsilon^p(\kappa(s)) m_{2p}^{1/2} \mathbf{P}_f[\Lambda_{\hat{k}}] + (2S_\varepsilon(\kappa(s)))^p \,.$$

Moreover we have:

$$\begin{aligned}
\mathbf{P}_f[\Lambda_{\hat{k}}] &\leq \mathbf{P}_f\left[\bigcup_{k\in\mathcal{Z}_\varepsilon} \Lambda_k\right] \\
&\leq \sum_{k\in\mathcal{Z}_\varepsilon} 2^{-|k|} \\
&\leq \sum_{n=0}^{\infty} \sum_{k\in\mathcal{Z}_n} 2^{-n} \\
&\leq \sum_{n=0}^{\infty} (\#\mathcal{Z}_n) 2^{-n} < +\infty.
\end{aligned}$$

Let us denote $|\mathcal{Z}| = \sum_{n=0}^{\infty} (\#\mathcal{Z}(n)) 2^{-n}$. We obtain:

$$(*) \leq \sigma_\varepsilon^p(\kappa(s)) m_{2p}^{1/2} |\mathcal{Z}| + (2S_\varepsilon(\kappa(s)))^p \leq \left(2^p + m_{2p}^{1/2}|\mathcal{Z}|\right) S_\varepsilon^p(\kappa(s)).$$

It is not difficult to obtain a similar result for the last expectation:

$$\mathbf{E}_f\left[\left(\sigma_\varepsilon(\hat{k})|\xi(\hat{k})|\right)^p \mathbf{1}_{\{\hat{k}\in B_1(s)\}}\right] \leq \left(2^p + m_{2p}^{1/2}|\mathcal{Z}|\right) S_\varepsilon^p(\kappa(s)).$$

Finaly, using (2.10) and the control of the expaectations we obtain:

$$R(s,p) \leq C_p S_\varepsilon^p(\kappa(s)) \tag{2.11}$$

where C_p is a constant depending only on p and $|\mathcal{Z}|$.

Back to our problem. Now, we can conclude. Let us recall that:

$$(**) = \mathbf{E}_f\left[\left(\left|f_\varepsilon^\Phi(t) - f(t)\right|\right)^q\right] \leq R(0,q) + \sum_{s\geq 1} \sqrt{R(s,2q)D(s)}.$$

Thus, we obtain — using (2.6) and (2.11):

$$(**) \leq C_q S_\varepsilon^q(\kappa) + (2C_{2q})^{1/2} \sum_{s\geq 1} \sqrt{S_\varepsilon^{2q}(\kappa(s)) \sum_{l\in\mathcal{Z}_\varepsilon, l\succeq\kappa(s-1)} 2^{-\tilde{C}|l|}}.$$

Let us recall that $\tilde{C} = 3q + 2$ and that:

$$S_\varepsilon(\kappa(s)) = S_\varepsilon(0) 2^{|\kappa(s)|} \sqrt{1 + |\kappa(s)| \ln 2}.$$

Thus:

$$\begin{aligned} S_\varepsilon^{2q}(\kappa(s)) 2^{-3q|\kappa(s-1)|} &\leq S_\varepsilon^{2q}(0) 2^{2q|\kappa(s)|} (1 + |\kappa(s)| \ln 2)^q 2^{-3q|\kappa(s)|} \\ &\leq S_\varepsilon^{2q}(0) 2^{2q(|\kappa(s)| - |\kappa(s-1)|)} \left(\frac{1 + |\kappa(s)| \ln 2}{2^{|\kappa(s-1)|}} \right)^q \\ &\leq 3^{dq} S_\varepsilon^{2q}(0). \end{aligned}$$

Now, it is easy to see that (we do not recall that $l \in \mathcal{Z}_\varepsilon$):

$$\begin{aligned} (**) &\leq C_q S_\varepsilon^q(\kappa) + (3^{dq} 2 C_{2q})^{1/2} S_\varepsilon^q(0) \sum_{s \geq 1} \sqrt{\sum_{l \succeq \kappa(s-1)} 2^{-3q(|l| - |\kappa(s-1))| - 2|l|}} \\ &\leq C_q S_\varepsilon^q(\kappa) + (3^{dq} 2 C_{2q})^{1/2} S_\varepsilon^q(0) \sum_{s \geq 1} \sqrt{\sum_{l \succeq \kappa(s-1)} 2^{-2|l|}} \\ &\leq C_q S_\varepsilon^q(\kappa) + (3^{dq} 2 C_{2q})^{1/2} S_\varepsilon^q(0) \sum_{s \geq 1} \sum_{l \succeq \kappa(s-1)} 2^{-|l|} \end{aligned}$$

Now, we have to prove that:

$$\sum_{s \geq 1} \sum_{l \succeq \kappa(s-1)} 2^{-|l|} < +\infty.$$

Let us calculate:

$$\begin{aligned} \sum_{s \geq 1} \sum_{l \succeq \kappa(s-1)} 2^{-|l|} &= \sum_{s \geq 0} \sum_{n \geq |\kappa(s)|} (\#\mathcal{Z}(n)) \, 2^{-n} \\ &= \sum_{s \geq 0} \sum_{n \geq s} (\#\mathcal{Z}(n)) \, 2^{-n} \\ &= \sum_{n \geq 0} \sum_{s \leq n} (\#\mathcal{Z}(n)) \, 2^{-n} \\ &= \sum_{n \geq 0} \frac{n(n+1)}{2} (\#\mathcal{Z}(n)) \, 2^{-n} < +\infty. \end{aligned}$$

Let us denote $\|\mathcal{Z}\|$ this constant. Finaly, if we remeber that $\kappa = k(\beta, L, \varepsilon)$ and that $S_\varepsilon(0) \leq S_\varepsilon(\kappa)$, we obtain the following result:

$$\mathbf{E}_f \left[\left(\left| f_\varepsilon^\Phi(t) - f(t) \right| \right)^q \right] \leq \left(C_q + (3^{dq} 2 C_{2q})^{1/2} \|\mathcal{Z}\| \right) S_\varepsilon(k(\beta, L, \varepsilon)).$$

As $S_\varepsilon(k(\beta, L, \varepsilon)) = \varphi_\varepsilon(\beta, L)$, result follows.

B) $\mathcal{A}$ is empty

This case is simpler. We have to control:

$$\mathbf{E}_f\left[|f(t)|^q \mathbf{1}_{\{\mathcal{A}=\emptyset\}}\right] \leq L^q \mathbf{P}_f[\mathcal{A} = \emptyset]\,.$$

Moreover, we can assume that there exist $s \in \mathbf{N}$ such that $\kappa(s) = N_\varepsilon$ and $\kappa(s) \in \mathcal{Z}_\varepsilon$. The fact that $\mathcal{A}$ is empty implies that $\kappa(s)$ is not in $\mathcal{A}$, thus:

$$\mathbf{P}_f[\mathcal{A} = \emptyset] \leq \mathbf{P}_f[\kappa(s) \notin \mathcal{A}]\,.$$

The same quantity was controlled by formula (2.6). Then, it is easy to obtain that:

$$\mathbf{P}_f[\mathcal{A} = \emptyset] \leq \left(\sum_{n\geq 0} (\#\mathcal{Z}_{2n})\, 2^{-\tilde{C}n+1}\right) 2^{-\tilde{C}|\kappa(s)|}.$$

The last thing we have to observe is that $2^{-\tilde{C}|\kappa(s)|} \leq S_\varepsilon(\kappa)$.

2.7 Proof of theorem 2

Let us consider another admissible family $\Psi = \{\psi_\varepsilon(\beta, L)\}_{(\beta,L)\in\mathcal{B}\times\mathcal{I}}$ and f_ε^Ψ an estimator satisfying (A.U.B) with Ψ.

To prove that Φ is the adaptive rate, it is enough to prove the following assertion:

Lemma 7. *Set $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathcal{B}$ and $\beta = (\beta_1, \ldots, \beta_d) \in \mathcal{B}$ such that $\bar{\alpha} < \bar{\beta}$. Set L_α and L_β in $\mathcal{I}$. If*

$$\frac{\psi_\varepsilon(\alpha, L_\alpha)}{\varphi_\varepsilon(\alpha, L_\alpha)} \xrightarrow[\varepsilon\to 0]{} 0,$$

then,

$$\frac{\psi_\varepsilon(\beta, L_\beta)}{\varphi_\varepsilon(\beta, L_\beta)} \times \frac{\psi_\varepsilon(\alpha, L_\alpha)}{\varphi_\varepsilon(\alpha, L_\alpha)} \xrightarrow[\varepsilon\to 0]{} +\infty.$$

Indeed, let us remark, first, that we cannot improve $\varphi_\varepsilon(b, L)$ because it is the minimax rate of convergence on $H(b, L)$ for all L.

Next, let us suppose that there exists (β_0, L_0) such that $\psi_\varepsilon(\beta_0, L_0)$ improves $\varphi_\varepsilon(\beta_0, L_0)$. Using the previous lemma, it is easy to sea that $\mathcal{I}_0(\Psi/\Phi) \subset$

$\mathcal{B}(\bar{\beta}_0) \times \mathcal{I}$. Indeed, let us suppose that there exists $(\beta_1, L1)$ such that $\psi_\varepsilon(\beta_1, L_1)$ improves $\varphi_\varepsilon(\beta_1, L_1)$ and $\bar{\beta}_1 < \bar{\beta}_0$ then we obtain:

$$\frac{\psi_\varepsilon(\beta_1, L_{\beta_1})}{\varphi_\varepsilon(\beta_1, L_{\beta_1})} \times \frac{\psi_\varepsilon(\beta_0, L_{\beta_0})}{\varphi_\varepsilon(\beta_0, L_{\beta_0})} \underset{\varepsilon \to 0}{\longrightarrow} +\infty.$$

In particular $\psi_\varepsilon(\beta_0, L_{\beta_0})/\varphi_\varepsilon(\beta_0, L_{\beta_0})$ tends to $+\infty$ which it is impossible.

On the other hand, it is easy to sea that $\bigcup_{\gamma > \beta_0} \mathcal{B}(\gamma) \times \mathcal{I} \subset \mathcal{I}_\infty(\Psi/\Phi)$.

Lemma 7 is a corrolary of the following proposition:

Proposition 1. *Set $(\alpha, \beta) \in \mathcal{B}^2$ such that $\bar{\alpha} < \bar{\beta}$ and $(L_\alpha, L_\beta) \in \mathcal{I}^2$. Let us define, for any estimator $\tilde{f}_\varepsilon(\cdot)$ which satisfies (A.U.B.) with Ψ and for all $\nu < 2(\bar{\beta} - \bar{\alpha})/((2\bar{\alpha}+1)(2\bar{\beta}+1))$ the following quantiy:*

$$\begin{aligned} R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) &= \sup_{f \in H(\alpha, L_\alpha)} \mathbf{E}_f\Big[\Big(\varphi_\varepsilon^{-1}(\alpha, L_\alpha)|\tilde{f}_\varepsilon(t) - f(t)|\Big)^q\Big] \\ &\quad + \sup_{f \in H(\beta, L_\beta)} \mathbf{E}_f\Big[\Big(\varepsilon^\nu \varphi_\varepsilon^{-1}(\beta, L_\beta)|\tilde{f}_\varepsilon(t) - f(t)|\Big)^q\Big]. \end{aligned}$$

If $\psi_\varepsilon(\alpha, L_\alpha)/\varphi_\varepsilon(\alpha, L_\alpha)$ tends to 0 as $\varepsilon \to 0$, then, we have:

$$\liminf_{\varepsilon \to 0} R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) > 0.$$

Proof. As $\psi_\varepsilon(\delta, L_\delta)$ and $\varphi_\varepsilon(\delta, L_\delta)$ do not depend, in order, on L_δ ($\delta \in \{\alpha, \beta\}$), we will denote, for simplicity:

$$\psi_\varepsilon(\delta) \triangleq \psi_\varepsilon(\delta, L_\delta) \text{ and } \varphi_\varepsilon(\delta) \triangleq \Big(\varepsilon\sqrt{\ln 1/\varepsilon}\Big)^{\frac{2\bar{\delta}}{2\bar{\delta}+1}}.$$

Set $\varkappa$ a positive parameter to be chosen. We consider $h_i = h_i(\varepsilon)$ defined by the formula

$$h_i = \Big(\varkappa\varepsilon\sqrt{\ln \varepsilon^{-1}}\Big)^{\frac{2\bar{\alpha}}{2\bar{\alpha}+1}\frac{1}{\alpha_i}},$$

and we consider the two following functions:

$$\left\{ \begin{array}{rcl} f_0 & = & 0 \\ f_1(x) & = & L_\alpha \varkappa^{\frac{2\bar{\alpha}}{2\bar{\alpha}+1}} \varphi_\varepsilon(\alpha) f\left(\frac{x_1 - t_1}{h_1}, \ldots, \frac{x_d - t_d}{h_d}\right) \end{array} \right.$$

where f belongs to $H(\alpha, 1)$. Hence f_1 belongs to $H(\alpha, L_\alpha)$ and, if $\mathbf{E}_0$ and $\mathbf{E}_1$ denote respectively $\mathbf{E}_{f_0}$ and $\mathbf{E}_{f_1}$, we have:

$$\begin{aligned} R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) &\geq \mathbf{E}_0\Big|\varepsilon^\nu \varphi_\varepsilon^{-1}(\beta)\tilde{f}_\varepsilon(t)\Big|^q + \mathbf{E}_1\Big|\varphi_\varepsilon^{-1}(\alpha)(\tilde{f}_\varepsilon(t) - f_1(t))\Big|^q \\ &\geq \mathbf{E}_0\Big|\varepsilon^\nu \varphi_\varepsilon^{-1}(\beta)\tilde{f}_\varepsilon(t)\Big|^q + \mathbf{E}_1\Big|\varphi_\varepsilon^{-1}(\alpha)\tilde{f}_\varepsilon(t) - z\Big|^q, \end{aligned}$$

where z denote $L\varkappa^{\frac{2\bar{\alpha}}{2\bar{\alpha}+1}}f(0)$. For simplicity, we consider the following notations:

$$\lambda_\varepsilon = \varepsilon^\nu \frac{\varphi_\varepsilon(\alpha)}{\varphi_\varepsilon(\beta)} = \varepsilon^\nu \left(\varepsilon\sqrt{\ln\frac{1}{\varepsilon}}\right)^{-\varrho} \text{ where } \varrho = \frac{2(\bar{\beta}-\bar{\alpha})}{(2\bar{\beta}+1)(2\bar{\alpha}+1)}.$$

and

$$\tilde{\theta} = \varphi_\varepsilon^{-1}(\alpha)|\tilde{f}_\varepsilon(t)|.$$

Thus, we have:

$$R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) \geq \mathbf{E}_0\left|\lambda_\varepsilon\tilde{\theta}\right|^q + \mathbf{E}_1\left|\tilde{\theta}-z\right|^q$$

By changing the probability measure, we obtain:

$$R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) \geq \mathbf{E}_1\left[\left|\lambda_\varepsilon\tilde{\theta}\right|^q Z_\varepsilon + \left|\tilde{\theta}-z\right|^q\right]$$

where Z_ε denotes the classical likelyhood ratio:

$$Z_\varepsilon = \frac{d\mathbf{P}_0}{d\mathbf{P}_1}(\mathcal{X}^{(\varepsilon)}).$$

Now, consider the following event for $\delta > 0$:

$$\Lambda = \left\{|\tilde{\theta}| > \delta\right\}$$

Clearly, if δ is small enough:

$$R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) \geq \mathbf{E}_1\left[(\delta\lambda_\varepsilon)^q Z_\varepsilon \mathbf{1}_{\{\Lambda\}} + (z-\delta)^q \mathbf{1}_{\{\Lambda^c\}}\right].$$

But,

$$\begin{aligned} Z_\varepsilon &= \exp\left(-\frac{1}{\varepsilon}\int_{\mathbf{R}^d} f_1(u)dW(u) - \frac{1}{2\varepsilon^2}\|f_1\|^2\right) \\ &= \exp\left(-\frac{\|f_1\|}{\varepsilon}\xi - \frac{1}{2}\left(\frac{\|f_1\|}{\varepsilon}\right)^2\right), \end{aligned}$$

where $\xi \sim \mathcal{N}(0,1)$. Hence, if the event

$$\Lambda^a = \{|\xi| \leq a\}$$

occurs, we can deduce that:

$$\begin{aligned} Z_\varepsilon &\geq \exp\left(-\frac{\|f_1\|}{\varepsilon}a - \frac{1}{2}\frac{\|f_1\|^2}{\varepsilon^2}\right) \\ &\geq \exp\left(-\frac{1}{2}\left(\frac{\|f_1\|}{\varepsilon}+a\right)^2\right) \end{aligned}$$

Then, we have:

$$\begin{aligned} R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) &\geq \mathbf{E}_1\left[(\delta\lambda_\varepsilon)^q Z_\varepsilon \mathbf{1}_{\{\Lambda\cap\Lambda^a\}} + (z-\delta)^q \mathbf{1}_{\{\Lambda^c\}}\right] \\ &\geq \mathbf{E}_1\left[(\delta\lambda_\varepsilon)^q \exp\left(-\frac{1}{2}\left(\frac{\|f_1\|}{\varepsilon}+a\right)^2\right)\mathbf{1}_{\{\Lambda\cap\Lambda^a\}} + (z-\delta)^q \mathbf{1}_{\{\Lambda^c\}}\right]. \end{aligned}$$

Let us remark that:

$$\frac{\|f_1\|}{\varepsilon} = L\varkappa\|f\|\sqrt{\ln\frac{1}{\varepsilon}}.$$

If we chose

$$a = L\varkappa\|f\|\sqrt{\ln\frac{1}{\varepsilon}} \wedge 1,$$

then we obtain:

$$\begin{aligned} R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) &\geq \mathbf{E}_1\left[(\delta\lambda_\varepsilon)^q \exp\left(-\left(\frac{\|f_1\|}{\varepsilon}\right)^2\right)\mathbf{1}_{\{\Lambda\cap\Lambda^a\}} + (z-\delta)^q \mathbf{1}_{\{\Lambda^c\}}\right] \\ &\geq \mathbf{E}_1\left[(\delta\lambda_\varepsilon)^q \mathcal{E}^{(L_\alpha\varkappa\|f\|)^2}\mathbf{1}_{\{\Lambda\cap\Lambda^a\}} + (z-\delta)^q \mathbf{1}_{\{\Lambda^c\}}\right] \\ &\geq \mathbf{E}_1\left[(\delta\eta_\varepsilon)^q \mathcal{E}^{q(\nu-\varrho)+(L_\alpha\varkappa\|f\|)^2}\mathbf{1}_{\{\Lambda\cap\Lambda^a\}} + (z-\delta)^q \mathbf{1}_{\{\Lambda^c\}}\right], \end{aligned}$$

where $\eta_\varepsilon = \left(\ln\frac{1}{\varepsilon}\right)^{-\varrho/2}$.

Let us introduce

$$t_\varepsilon = \frac{q}{\ln\frac{1}{\varepsilon}}\left(\ln\frac{1}{\delta\eta_\varepsilon} + \ln AL_\alpha f(0)\right) \to 0,$$

where

$$A = \left(\frac{\sqrt{q(\varrho-\nu)}}{L_\alpha\|f\|}\right)^{\frac{2\bar{\alpha}}{2\bar{\alpha}+1}},$$

and let us chose:

$$\varkappa = \frac{\sqrt{q(\varrho-\nu)-t_\varepsilon}}{L_\alpha\|f\|}.$$

Using this choice of $\varkappa$, we obtain that:

$$(L_\alpha\varkappa\|f\|)^2 = q\varrho - t_\varepsilon \text{ and } (\delta\eta_\varepsilon)^q \mathcal{E}^{-q\varrho+(L_\alpha\varkappa\|f\|)} = (AL_\alpha f(0))^q$$

Thus, we obtain:

$$\begin{aligned} R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) &\geq \mathbf{E}_1\left[(AL_\alpha f(0))^q \mathbf{1}_{\{\Lambda\cap\Lambda^a\}} + (\varkappa^{\frac{2\bar{\alpha}}{2\bar{\alpha}+1}} L_\alpha f(0) - \delta)^q \mathbf{1}_{\{\Lambda^c\}}\right] \\ &> \mathbf{E}_1\left[(\varkappa^{\frac{2\bar{\alpha}}{2\bar{\alpha}+1}} L_\alpha f(0) - \delta)^q \left(\mathbf{1}_{\{\Lambda+\Lambda^c\}}\right)\mathbf{1}_{\{\Lambda^a\}}\right] \\ &\geq (\varkappa^{\frac{2\bar{\alpha}}{2\bar{\alpha}+1}} L_\alpha f(0) - \delta)^q \mathbf{P}_1\left[\Lambda^a\right] \\ &\geq (\varkappa^{\frac{2\bar{\alpha}}{2\bar{\alpha}+1}} L_\alpha f(0) - \delta)^q \mathbf{P}\left[|\xi| \geq 1\right]. \end{aligned}$$

And, then:

$$\liminf_{\varepsilon\to 0} R_\varepsilon^{(q)}(\tilde{f}_\varepsilon) \geq \left(L_\alpha^{\frac{1}{2\bar{\alpha}+1}} \frac{f(0)}{\|f\|^{\frac{2\bar{\alpha}}{2\bar{\alpha}+1}}} \left(q(\varrho-\nu)\right)^{\frac{\bar{\alpha}}{2\bar{\alpha}+1}} \right)^q \mathbf{P}\left[|\xi| \geq 1\right] > 0.$$

Proposition is proved.

Chapter 3

Minimax Procedure — Partially Case

3.1 Introduction

3.1.1 Statistical model

This paper is the second part of our paper "Fully adaptive case". Further, we will refer to this paper as (Part I). We consider the same model. Our observations $\mathcal{X}^{(\varepsilon)} = (X_\varepsilon(u))_{u\in[0,1]^d}$ satisfies the same SDE:

$$X_\varepsilon(du) = f(u)du + \varepsilon W(du), \quad \forall u \in [0,1]^d,$$

where $f : \mathbf{R}^d \to \mathbf{R}$ is an unknown signal to be estimated, W is a standartd Gaussian white noise from $\mathbf{R}^d$ to $\mathbf{R}$ and ε is the noise level.

Our main goal is to estimate f at a fixed point $t \in (0,1)^d$.

3.1.2 Our goal

In this second part of our article, we study the "Partially adaptive case". Let us recall that we are interrested in pointwise estimation among the class of anisotropic Hölder spaces. Let us recall some notations: $l_* < l^*$ and

$b = (b_1, \dots, b_d)$ are given. Moreover, we consider only Hölder spaces $H(\beta, L)$ (defined in Part I) such that

$$\beta \in \mathcal{B} = \prod_{i=1}^{d} (0; b_i] \text{ and } L \in \mathcal{I} = [l_*; l^*].$$

Remark 11. *Let us just recall that $\beta = (\beta_1, \dots, \beta_d)$ can be viewed as the smoothness parameter. Each β_i represents the smoothness of a function in direction i. Moreover, L is a Lipschitz constant.*

We denote $\Sigma = \bigcup_{\mathcal{B}\times\mathcal{I}} H(\beta, L)$. Our goal is to answer this questions: Is it possible to guarantee a quality of estimation? On which space (included in Σ)? With which procedure of estimation?

For example, if we consider $\tilde{\eta}_\varepsilon(\gamma) = \varepsilon^{2\gamma/(2\gamma+1)}$, it is well known that we can guarantee this quality on each space $H(\beta, L)$ such that $\bar{\beta} = \gamma$ (because it is the minimax rate of convergence on this space) using the minimax on this space estimator. But one of our results implies that we cannot guarantee this quality simultaneously on each such space.

Now, we fix $0 < \gamma < \bar{b}$, and we consider

$$\eta_\varepsilon(\gamma) = (l^*)^{\frac{1}{2\gamma+1}} \left(\|K\| \varepsilon \sqrt{\ln \ln \frac{1}{\varepsilon}} \right)^{\frac{2\gamma}{2\gamma+1}}.$$

Our result is that there exists an estimator, namely $f_\varepsilon^\gamma(\cdot)$, such that $\eta_\varepsilon(\gamma)$ is the minimax rate of convergence of this estimator on $\Sigma(\gamma)$ defined by

$$\Sigma(\gamma) = \bigcup_{(\beta,L)\in\mathcal{B}(\gamma)\times\mathcal{I}} H(\beta, L) = \bigcup_{\beta\in\mathcal{B}(\gamma)} H(\beta, l^*)$$

where

$$\mathcal{B}(\gamma) = \left\{ \beta \in \mathcal{B} : \bar{\beta} = \gamma \right\}.$$

Thus, using f_ε^γ as procedure of estimation, we can guarantee that the quality is $\eta_\varepsilon(\gamma)$, at least on $\Sigma(\gamma)$.

3.1.3 Result

Theorem 3. *Our result consists in two inequalities:*

$$\limsup_{\varepsilon\to 0} \sup_{f\in\Sigma(\gamma)} \mathbf{E}_f \left[\left(\eta_\varepsilon(\gamma)^{-1} |f_\varepsilon^\gamma(t) - f(t)| \right)^q \right] < +\infty. \qquad \text{(U.B.)}$$

$$\liminf_{\varepsilon\to 0}\inf_{\tilde f}\sup_{f\in\Sigma(\gamma)}\mathbf{E}_f\left[\left(\eta_\varepsilon(\gamma)^{-1}|\tilde f(t)-f(t)|\right)^q\right]>0, \qquad \text{(L.B.)}$$

where the infimum is taken over all possible estimators.

In words, f_ε^γ is a minimax on $\Sigma(\gamma)$ estimator.

This paper consists in the proof of this assertion. First, we construct the estimator f_ε^γ. Next, we prove the corresponding lower bound.

Remark 12. *Let us remark that this result can be viewed as an adaptive result. Indeed, let us consider $\Sigma(\gamma)$ as a family —instead of an union— of Hölder spaces $H(\beta,L)$ such that $\bar\beta=\gamma$. It is well known that on each $H(\beta,L)$ there exists a minimax on this space estimator which depends explicitely on (β,L) at least trough its bandwidth. Thus, question of adaptation arrizes naturally.*

Our lower bound proves that an optimal adaptive estimator $f^*(\cdot)$ such that

$$\limsup_{\varepsilon\to 0}\sup_{(\beta,L)\in\mathcal{B}(\gamma)\times\mathcal{I}}\sup_{f\in H(\beta,L)}\mathbf{E}_f\left[\left(\varepsilon^{-\frac{2\gamma}{2\gamma+1}}|f^*(t)-f(t)|\right)^q\right]<+\infty$$

does not exist.

Our upper bound proves that $f_\varepsilon^\gamma(\cdot)$ is an adaptive estimator. Moreover the price to pay is only $\sqrt{\ln\ln 1/\varepsilon}$ which is to be compared with the classical loss $\sqrt{\ln 1/\varepsilon}$ in other adaptive problems.

Moreover we prove that our estimator is optimal in a minimax sense.

3.2 Procedure

3.2.1 Collection of kernel estimators

Let us recall that kernels were defined in the first part of this paper: "fully adaptive case". Here we have just to chose a good collection of kernel estimators.

Let us define

$$n_\varepsilon(\gamma)=\left\lfloor\frac{1}{\ln 2}\left(\frac{4(\bar b-\gamma)}{(2\bar b+1)(2\gamma+1)}\ln\frac{l^*}{\|K\|\varepsilon}-\frac{1}{2\gamma+1}\ln\ln\ln\frac{1}{\varepsilon}\right)\right\rfloor.$$

Let us denote

$$\mathcal{Z}_\gamma^\varepsilon = \mathcal{Z}(n_\varepsilon(\gamma)).$$

Let us recall the definition of this set:

$$\mathcal{Z}(n) = \left\{ k =\in \mathbf{Z}^d : \sum_{i=1}^d (k_i + 1) = n \text{ and } \forall i, |k_i| \leq C(b)n + 1 \right\},$$

where

$$C(b) = \frac{2\bar{b}+1}{2\bar{b}} \times \frac{\ln 2 + \sqrt{2\ln 2}}{\ln 2}.$$

Finally, we consider the following collection $\{\hat{f}_k(\cdot)\}_{k\in\mathcal{Z}_\gamma^\varepsilon}$.

3.2.2 Notations

Let us recall the following notation: for all $k \in \mathcal{Z}_\gamma^\varepsilon$, we have

$$\sigma_\varepsilon(k) = \frac{\varepsilon \|K\|}{\left(\prod_{i=1}^d h_i^{(k)}\right)^{1/2}},$$

where $h^{(k)} = (h_1^{(k)}, \dots, h_d^{(k)})$ is defined by:

$$h_i^{(k)} = (\|K\|\varepsilon)^{\frac{2\bar{b}}{2\bar{b}+1}\frac{1}{b_i}} 2^{-(k_i+1)}.$$

It is clear that for all k and l in $\mathcal{Z}_\gamma^\varepsilon$, $\sigma_\varepsilon(k) = \sigma_\varepsilon(l) \triangleq \sigma_\varepsilon(\gamma)$ and moreover that:

$$\sigma_\varepsilon(\gamma)\sqrt{\ln\ln\frac{1}{\varepsilon}} \asymp \eta_\varepsilon(\gamma).$$

Following the same strategy as in the first part of our paper, let us define the set $\mathcal{A}$ as follows: an index $k \in \mathcal{Z}_\gamma^\varepsilon$ belongs to $\mathcal{A}$ if it satisfies:

$$\left|\hat{f}_{k\wedge l}(t) - \hat{f}_l(t)\right| \leq C\sigma_\varepsilon(\gamma)\sqrt{\ln\ln\frac{1}{\varepsilon}}, \quad \forall l \neq k, l \in \mathcal{Z}_\gamma^\varepsilon,$$

where $k \wedge l$ denote the index $(k_i \wedge l_i)_{i=1,\dots,d}$.

3.2.3 Definition of our procedure

First of all, let us reformulate one of our result obtained in the first part of this paper: there exists an estimator, namely $f_\varepsilon^\Phi(\cdot)$, sucht that

$$\limsup_{\varepsilon\to 0}\sup_{\beta\in\mathcal{B}}\sup_{f\in H(\beta,l^*)}\mathbf{E}_f\left[\left(\left(\varepsilon\sqrt{\ln\frac{1}{\varepsilon}}\right)^{-\frac{2\bar{\beta}}{2\bar{\beta}+1}}|f_\varepsilon^\Phi(t)-f(t)|\right)^{2q}\right]<+\infty.$$

Now, let us define our new estimator: If the random set $\mathcal{A}$ is non-empty, we chose arbitrary any index which belongs to this set. We denote $\hat{k}$ a such index. Then we construct:

$$f_\varepsilon^\gamma(\cdot)=\hat{f}_{\hat{k}}(\cdot).$$

On the other hand, if $\mathcal{A}$ is empty, we define

$$f_\varepsilon^\gamma(\cdot)=f_\varepsilon^\Phi(\cdot).$$

Remark 13. *This procedure is closed to the adaptive one. The main difference consists in the following: when the set $\mathcal{A}$ is empty, we estimate using a best estimator than 0. In fact the probability $\mathbf{P}_f[\mathcal{A}=\emptyset]$ is too large to use a trivial estimator.*

3.3 Proof of (U.B)

3.3.1 Method

First of all, let us recall that our minimax on Σ_γ estimator is in fact an "adaptive procedure of estimation" because the real smoothness parameter is unknown.

Thus, the mechanism of the proof is very closed to the previous one. We will compare the estimator chosen by our procedure with respect to the "best" estimator among our class but depending on the unknown parameter.

First, we have to define correctly all indexes we need. Next, we will be able to prove the result. Moreover, as the class Σ_γ depends only in L though l^* (because $H(\beta,L)\subset H(\beta,l^*)$), we will assume that $l_*=l^*=1$ to make the proof simpler. Consequently we will denote $H(\beta)$ instead of $H(\beta,1)$.

3.3.2 Indexes

Let us suppose that our unknown signal in Σ_γ belongs to $H(\beta)$ with $\bar{\beta} = \gamma$. Clearly, if we consider the kernel estimator defined using bandwidth

$$\left(h_i(\beta, \varepsilon) = \left(\|K\| \varepsilon \sqrt{\ln \ln \frac{1}{\varepsilon}} \right)^{\frac{2\gamma}{2\gamma+1}\frac{1}{\beta_i}} \right)_{i=1,\dots,d},$$

it achieves the expected rate $\eta_\varepsilon(\gamma)$.

We consider the bandwidth

$$h^*(\varepsilon) = \left(h_i^*(\varepsilon) = (\|K\|\varepsilon)^{\frac{2\bar{b}}{2\bar{b}+1}\frac{1}{b_i}} \right)_{i=1,\dots,d}$$

and define the following indexes: for all $i \in [\![1; d]\!]$ and $\beta \in \mathcal{B}$ such that $\bar{b} = \gamma$, we construct

$$\tilde{k}_i(\beta, \varepsilon) = \left\lfloor \frac{1}{\ln 2} \ln \frac{h_i^*(\varepsilon)}{h_i(\beta, \varepsilon)} \right\rfloor.$$

If $h^{(k)} = (h_i^{(k)})_i$ denote the bandwidth defined by

$$h_i^{(k)} = h_i^*(\varepsilon) 2^{-(k_i+1)},$$

we obtain clearly that the kernel estimator defined using bandwidth $h^{(\tilde{k}(\beta,\varepsilon))}$ is asymptotically as good as that one defined using $h(\beta, \varepsilon)$.

Now, let us define:

$$k_i(\beta, \varepsilon) = \begin{cases} \tilde{k}_i(\beta, \varepsilon) & \text{if } i = 1, \dots, d-1 \\ n_\varepsilon(\gamma) - 1 - \sum_{i=1}^{d_1} (\tilde{k}_i(\beta, \varepsilon) + 1) & \text{otherwise} \end{cases}$$

It is easy to prove that

$$\left| \tilde{k}_d(\beta, \varepsilon) - k_d(\beta, \varepsilon) \right| \leq d.$$

Thus, asymptotically, estimator defined using $h^{(k(\beta,\varepsilon))}$ is as good as that one defined using $h^{(\tilde{k}(\beta,\varepsilon))}$ and thus as good as that one defined by $h(\beta, \varepsilon)$.

Moreover it is simple, by producing similar arguments than in the first part of this paper, to obtain that $k(\beta, \varepsilon)$ belongs to $\mathcal{Z}(n_\varepsilon(\gamma))$.

3.3.3 Proof

We want to prove that, for all $\varepsilon < 1$:

$$\sup_{\beta\in\mathcal{B}(\gamma)}\sup_{f\in H(\beta)}\mathbf{E}_f\left[\left(\eta_\varepsilon^{-1}(\gamma)\,|f_\varepsilon^\gamma(t)-f(t)|\right)^q\right] < M_q(\gamma)$$

where $M_q(\gamma)$ is an explicit constant given in the proof.

Set $\varepsilon < 1$ and $\beta \in \mathcal{B}$ such that $\bar{\beta} = \gamma$. Let us suppose that $f \in H(\beta) \subset \Sigma_\gamma$. Set q a fixed parameter.

First, let us suppose that $\mathcal{A}$ is non empty.

A) $\mathcal{A}$ is non empty

Let us denote $\kappa = k(\beta, \varepsilon)$. Our goal is to majorate the following quatiy:

$$(*) = \mathbf{E}_f\left[|\hat{f}_{\hat{k}}(t) - f(t)|^q\right].$$

Let us consider

$$\begin{cases} I_1 &= |\hat{f}_{\hat{k}}(t) - \hat{f}_{\hat{k}\wedge\kappa}| \\ I_2 &= |\hat{f}_{\hat{k}\wedge\kappa}(t) - \hat{f}_{\kappa}(t)| \\ I_3 &= |\hat{f}_{\kappa}(t) - f(t)| \end{cases}$$

Let us remark that, if $\hat{k} = \kappa$, then $I_1 = I_2 = 0$. Thus we can suppose that $\hat{k} \neq \kappa$.

a) Let us control of $\mathbf{E}_f[I_3^q]$. Using lemma **??**, we have:

$$\begin{aligned} \mathbf{E}_f[I_3^q] &= \mathbf{E}_f\left[|\hat{f}_{\kappa}(t) - f(t)|^q\right] \\ &\leq \mathbf{E}_f[(|b_\kappa(t) - f(t)| + \sigma_\varepsilon(\gamma)|\xi(\kappa)|)^q] \\ &\leq \mathbf{E}_f\left[\left(B^\beta(\kappa) + \sigma_\varepsilon(\gamma)|\xi(\kappa)|\right)^q\right] \\ &\leq \mathbf{E}_f[(C^* S_\varepsilon(\kappa) + \sigma_\varepsilon(\gamma)|\xi(\kappa)|)^q] \\ &\leq \left(\sigma_\varepsilon(\gamma)\sqrt{\ln\ln\frac{1}{\varepsilon}}\right)^q \mathbf{E}_f\left[\left(C^* + \frac{|\xi(\kappa)|}{\sqrt{\ln\ln\frac{1}{\varepsilon}}}\right)^q\right] \end{aligned}$$

b) Let us control $\mathbf{E}_f[I_2^q]$. Our procedure control itself this expectation. We have:

$$\mathbf{E}_f[I_2^q] \leq C^q\left(\sigma_\varepsilon(\gamma)\sqrt{\ln\ln\frac{1}{\varepsilon}}\right)^q.$$

Let us remark that we use the fact that κ belongs to $\mathcal{Z}_\gamma^\varepsilon$.

c) Finally, let us control $\mathbf{E}_f[I_1^q]$. Using lemma **??**, we ontain:

$$\begin{aligned}
\mathbf{E}_f[I_1^q] &= \mathbf{E}_f\Big[|\hat{f}_{\hat{k}}(t) - \hat{f}_{\hat{k}\wedge\kappa}|^q\Big] \\
&\leq \mathbf{E}_f\Big[\Big(2C^* S_\varepsilon(\kappa) + \sigma_\varepsilon(\hat{k})|\xi(\hat{k})| + \sigma_\varepsilon(\hat{k}\wedge\kappa)|\xi(\hat{k}\wedge\kappa)|\Big)^q\Big] \\
&\leq S_\varepsilon(\kappa)^q \mathbf{E}_f\left[\left(2C^* + \frac{|\xi(\hat{k})| + |\xi(\hat{k}\wedge\kappa)|}{\sqrt{\ln\ln\frac{1}{\varepsilon}}}\right)^q\right] \\
&\leq \mathbf{E}_f\left[\left(2C^* + \frac{|\xi(\hat{k})| + |\xi(\hat{k}\wedge\kappa)|}{\sqrt{\ln\ln\frac{1}{\varepsilon}}}\right)^q\right]\left(\sigma_\varepsilon(\gamma)\sqrt{\ln\ln\frac{1}{\varepsilon}}\right)^q
\end{aligned}$$

Finally, we obtain the following inequelity:

$$\begin{aligned}
(*) &\leq (3^{q-1}\vee 1)\left(\mathbf{E}_f[I_1^q] + \mathbf{E}_f[I_2^q] + \mathbf{E}_f[I_3^q]\right) \\
&\leq (3^{q-1}\vee 1)\left\{C^q + (2^q+1)(C^*)^q + o(1/\varepsilon)\right\}\left(\sigma_\varepsilon(\gamma)\sqrt{\ln\ln\frac{1}{\varepsilon}}\right)^q,
\end{aligned}$$

where $o(1/\varepsilon)$ tends to 0 where ε tends to 0. It is clear by applying Lebesgue's theorem.

B) $\mathcal{A}$ is empty

As $\mathcal{A}$ is empty, in particular κ does not belong to this set. Thus, we obtain:

$$\begin{aligned}
\mathbf{E}_f[|f_\varepsilon^\gamma(t) - f(t)|^q] &\leq \mathbf{E}_f\left[|f_\varepsilon^\Phi(t) - f(t)|^q \mathbf{1}_{\{\kappa\notin\mathcal{A}\}}\right] \\
&\leq \sqrt{\mathbf{E}_f[|f_\varepsilon^\Phi(t) - f(t)|^{2q}]\,\mathbf{P}_f[\kappa\notin\mathcal{A}]}
\end{aligned}$$

Using the upper bound of the first part of this paper we obtain:

$$\sqrt{\mathbf{E}_f[|f_\varepsilon^\Phi(t) - f(t)|^{2q}]} \leq \mathrm{Cte}\left(\varepsilon\sqrt{\ln\frac{1}{\varepsilon}}\right)^{\frac{2\gamma}{2\gamma+1}q}.$$

Thus, we have to control $\mathbf{P}_f[\kappa\notin\mathcal{A}]$. If $\kappa\notin\mathcal{A}$, there exists $l\in\mathcal{Z}_\varepsilon^\gamma$, $l\neq\kappa$, such that:

$$|\hat{f}_{\kappa\wedge l}(t) - \hat{f}_l(t)| > C\sigma_\varepsilon(\gamma)\sqrt{\ln\ln\frac{1}{\varepsilon}}.$$

And, consequently, we obtain that:

$$\begin{aligned}\mathbf{P}_f[\kappa \notin \mathcal{A}] &\leq \sum_{l \neq \kappa} \mathbf{P}_f\left[|\hat{f}_{\kappa\wedge l}(t) - \hat{f}_l(t)| > C\sigma_\varepsilon(\gamma)\sqrt{\ln\ln\frac{1}{\varepsilon}}\right] \\ &\leq 2\sum_{\kappa \neq l}\left(\frac{1}{\ln\frac{1}{\varepsilon}}\right)^{\frac{(C-2C^*)^2}{8}} \\ &\leq 2(\#\mathcal{Z}_\gamma^\varepsilon)\left(\frac{1}{\ln\frac{1}{\varepsilon}}\right)^{\frac{(C-2C^*)^2}{8}}.\end{aligned}$$

Moreover, it is easy to prove that there exists a constant C_b depending only on b such that:

$$\#\mathcal{Z}_\gamma^\varepsilon \leq C_b\left(\ln\frac{1}{\varepsilon}\right)^d.$$

On the other hand our choice of C implies that

$$\frac{(C-2C^*)^2}{8} = d + \frac{2\gamma}{2\gamma+1}(2q).$$

Thus, we obtain:

$$\mathbf{E}_f[|f_\varepsilon^\gamma(t) - f(t)|^q] \leq \text{Cte}\ \varepsilon^{\frac{2\gamma}{2\gamma+1}q}$$

3.4 Proof of (L.B.)

3.4.1 Method

The method is classical. Our goal is to minorate the minimax risk by a bayesian risk taken on a large number ($\sqrt{\ln 1/\varepsilon}$)of functions. In our mind, these functions are chosen because they represent the most difficult functions to be estimated in the considered class. This assertion is explained by lemma 9

3.4.2 Notations

Let us intoduce some basic notations. Let us fix $0 < \gamma < \bar{b}$. We say that a function $g : \mathbf{R}^d \to \mathbf{R}$ belongs to $\mathcal{G}(\gamma)$ if it satisfies:

$$\begin{cases} g(0) > 0. \\ \|g\| < +\infty \\ g \in \bigcap_{\beta \in \mathcal{B}(\gamma)} H(\beta) \\ \operatorname{supp} g \subset [-a; a]^d. \end{cases}$$

Here and later, we fix $g \in \mathcal{G}(\gamma)$.

Let us denote

$$\delta = \frac{\prod_{i=2}^d b_i}{\sum_{i=2}^d \prod_{j \neq i} b_j} = \frac{1}{1/\beta_2 + \cdots + 1/\beta_d}.$$

We consider

$$a = \left(\frac{1}{\gamma} - \frac{1}{\delta} \right)^{-1} < b_1,$$

and we denote $n_\varepsilon = \sqrt{\ln 1/\varepsilon}$.

Now, let us consider a family of vectors $\{\beta^{(k)}\}_k$ indexed by $k = 0, \ldots, n_\varepsilon$ and defined as follows:

$$\beta_1^{(k)} = a + k \frac{b_1 - a}{n_\varepsilon} \tag{3.1}$$

$$\beta_i^{(k)} = \frac{b_i}{\delta} \left(\frac{1}{\gamma} - \frac{1}{\beta_1^{(k)}} \right)^{-1} \quad \forall i = 2, \ldots, d. \tag{3.2}$$

Lemma 8. *For all $k = 0, \ldots, n_\varepsilon$ the vector $\beta^{(k)}$ belongs to $\mathcal{B}(\gamma)$.*

This lemma will be proved later.

Finally, let us introduce some functions. First of all, let us consider:

$$\forall i = 1, \ldots, d, \ \forall k = 0, \ldots, n_\varepsilon, \quad h_i^{(k)} = \left(\varkappa \varepsilon \sqrt{\ln \ln \frac{1}{\varepsilon}} \right)^{\frac{2\gamma}{2\gamma+1} \frac{1}{\beta_i^{(k)}}}$$

where $\varkappa < 1/(\sqrt{2}\|g\|)$. Then, we can define:

$$\begin{cases} f_0 \equiv 0 \\ f_k(x) = \varkappa^{\frac{2\gamma}{2\gamma+1}} \eta_\varepsilon(\gamma) g\left(\frac{x_1 - t_1}{h_1^{(k)}}, \ldots, \frac{x_d - t_d}{h_d^{(k)}} \right), \quad k \geq 1. \end{cases}$$

3.4.3 Proof

Now, let us prove our result. We will denote $\mathbf{P}_k$ instead of $\mathbf{P}_{f_k}$ and we consider the likelyhood ratio:

$$Z_\varepsilon = \frac{1}{n_\varepsilon} \sum_{k=1}^{n_\varepsilon} \frac{d\mathbf{P}_k}{d\mathbf{P}_0}(\mathcal{X}^{(\varepsilon)}).$$

This ratio satisfies the following lemma which will be proved further:

Lemma 9. *For all* $0 < \alpha < 1$*, we have:*

$$\limsup_{\varepsilon \to 0} \mathbf{P}_0[|Z_\varepsilon - 1| > \alpha] = 0.$$

Let us consider for any arbitrary estimator $\tilde{f}$, the following quantiy:

$$R_\varepsilon(\tilde{f}) = \sup_{f \in \Sigma_\gamma} \mathbf{E}_f \left[\left(\left(\varkappa \varepsilon \sqrt{\ln \ln \frac{1}{\varepsilon}} \right)^{-\frac{2\gamma}{2\gamma+1}} |\tilde{f}(t) - f(t)| \right)^q \right].$$

It is a well known result that, using bayesian method, for all $0 < \alpha < 1$ we obtain:

$$R_\varepsilon(\tilde{f}) \geq (1-\alpha) \left(\frac{g(0)}{2} \right)^q (1 - \mathbf{P}_0[|Z_\varepsilon - 1| > \alpha]).$$

Thus, we have:

$$\liminf_{\varepsilon \to 0} R_\varepsilon(\tilde{f}) \geq (1-\alpha) \left(\frac{g(0)}{2} \right)^q.$$

This inequality is equivalent to the following:

$$\liminf_{\varepsilon \to 0} \sup_{f \in \Sigma_\gamma} \mathbf{E}_f \left[\left(\eta_\varepsilon^{-1}(\gamma) |\tilde{f}(t) - f(t)| \right)^q \right] \geq (1-\alpha) \left(\varkappa^{\frac{2\gamma}{2\gamma+1}} \frac{g(0)}{2} \right)^q.$$

Now, if $\varkappa$ tends to $(\sqrt{2}\|g\|)^{-1}$ and α tends to 1 we obtain the lower bound:

$$\liminf_{\varepsilon \to 0} \sup_{f \in \Sigma_\gamma} \mathbf{E}_f \left[\left(\eta_\varepsilon^{-1}(\gamma) |\tilde{f}(t) - f(t)| \right)^q \right] \geq \left(2^{-(1+\gamma/(2\gamma+1))} \sup_{g \in \mathcal{G}(\gamma)} \frac{g(0)}{\|g\|} \right)^q.$$

Part III

Appendix

Appendix A

Fully Case

A.1 Proof of lemma 1

Let us denote $\mathcal{H} = (\mathbf{R}_+^*)^d$ and $\partial\mathcal{H} = \{h \in (\mathbf{R}_+^*)^d; \exists i \in [\![1; d]\!] h_i = 0\} \cup \{\infty\}$.

Let us recall that $\varphi_\varepsilon^{\beta,L} = b^{\beta,L} + s_\varepsilon^{\beta,L}$. It is easy to prove the following assertion:

$$\varphi_\varepsilon^{\beta,L}(h) \xrightarrow[h \to \partial\mathcal{H}]{} +\infty.$$

Thus, it is enough, to prove Lemma, to prove that $h(\beta, L, \varepsilon)$ is the unique point of $\mathcal{H}$ such that $\nabla\varphi_\varepsilon^{\beta,L} = 0$. Let us fix $i \in [\![1; d]\!]$, and let us calculate:

$$\partial_i\varphi_\varepsilon^{\beta,L}(h) = L\beta_i\lambda_i(\beta)h_i^{\beta_i-1} - \frac{\|K\|\varepsilon}{2\left(\prod_{j=1}^d h_j\right)^{\frac{1}{2}}}\frac{1}{h_i}\rho_\varepsilon(\beta, L).$$

To simplify notations, as β and L are fixed, we will denote λ_i instead of $\lambda_i(\beta)$. It follows that:

$$\partial_i\varphi_\varepsilon^{\beta,L}(h) = 0 \Leftrightarrow h_i^{\beta_i} = (\lambda_i\beta_i)^{-1}\frac{\|K\|}{2L}\frac{\varepsilon\rho_\varepsilon(\beta, L)}{\left(\prod_{j=1}^d h_j\right)^{\frac{1}{2}}}.$$

It is easy to deduce, from the previous equality, the following expression for h_i:

$$h_i = (\lambda_i\beta_i)^{\frac{-1}{\beta_i}}\left(\frac{\|K\|}{2L}\frac{\varepsilon\rho_\varepsilon(\beta, L)}{\left(\prod_{j=1}^d h_j\right)^{\frac{1}{2}}}\right)^{\frac{1}{\beta_i}}.$$

A simple computation prove that:

$$\left(\prod_{i=1}^{d} h_i\right)^{\frac{1}{2}} = \left(\prod_{i=1}^{d} (\lambda_i \beta_i)^{\frac{-1}{\beta_i}}\right)^{\frac{\bar{\beta}}{2\bar{\beta}+1}} \left(\frac{\|K\|}{2L}\varepsilon\rho_\varepsilon(\beta, L)\right)^{\frac{1}{2\bar{\beta}+1}}.$$

The follwing equality follows easily from previous equalities:

$$h_i = \gamma_i \left(\frac{\|K\|\Gamma}{2L}\varepsilon\rho_\varepsilon(\beta, L)\right)^{\frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i}}$$

where

$$\begin{cases} \gamma_i &= (\lambda_i\beta_i)^{\frac{-1}{\beta_i}} \\ \Gamma &= \left(\prod_{i=1}^{d}\gamma_i\right)^{\frac{1}{2}}. \end{cases}$$

Conversely, it is easy to proved that $h \in \mathcal{H}$ given by the previous formulas is such that $\nabla\varphi_\varepsilon^{\beta,L}(h) = 0$. This result implies that the problem is solved. Lemma is proved.

A.2 Proof of lemma 2

This lemma is the most technical one. It will be proved in two steps.

Step 1. *Set $(\beta, L) \in \mathcal{B} \times \mathcal{I}$. We have:*

$$0 \leq \sum_{i=1}^{d} (k_i(\beta, L, \varepsilon) + 1) \leq N_\varepsilon$$

where N_ε is defined by:

$$\left\lfloor 2\left(\frac{2\bar{b}}{2\bar{b}+1}\ln\frac{l_*}{\|K\|\varepsilon} + \ln\frac{l^*}{l_*}\right)\right\rfloor + 1.$$

Proof. Let us denote

$$\begin{aligned} x_\varepsilon(\beta, L) &= \frac{4(\bar{b}-\bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)}\ln\frac{L}{\|K\|\varepsilon} + \frac{2}{2\bar{b}+1}\ln\frac{L}{l_*} \\ &= \frac{4(\bar{b}-\bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)}\ln\frac{l_*}{\|K\|\varepsilon} + \frac{2}{2\bar{\beta}+1}\ln\frac{L}{l_*}. \end{aligned}$$

Using this notation, it is easy to proof that

$$\ln \prod_{i=1}^{d} \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} = x_\varepsilon(\beta, L) - \frac{1}{2\bar{\beta}+1} \ln(1 + x_\varepsilon(\beta, L)).$$

Moreover,

$$\frac{1}{2\bar{\beta}+1} \ln(1 + x_\varepsilon(\beta, L)) \leq \ln(1 + x_\varepsilon(\beta, L)) \leq x_\varepsilon(\beta, L),$$

thus

$$\ln \prod_{i=1}^{d} \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \geq 0.$$

This result implies that:

$$\sum_{i=1}^{d} (k_i(\beta, L, \varepsilon) + 1) \geq 0.$$

On the other hand,

$$\begin{aligned} \ln \prod_{i=1}^{d} \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} &\leq x_\varepsilon(\beta, L) \\ &= \frac{4(\bar{b}-\bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)} \ln \frac{l_*}{\|K\|\varepsilon} + \frac{2}{2\bar{\beta}+1} \ln \frac{L}{l_*} \\ &\leq \frac{4\bar{b}}{2\bar{b}+1} \ln \frac{l_*}{\|K\|\varepsilon} + 2 \ln \frac{l^*}{l_*} \end{aligned}$$

Step 2. *Set $(\beta, L) \in \mathcal{B} \times \mathcal{I}$ and let us denote $n = \sum_i (k_i(\beta, L, \varepsilon) + 1)$. Then, for all $i \in [\![1; d]\!]$, we have:*

$$|k_i(\beta, L, \varepsilon)| \leq \left(\frac{2\bar{b}+1}{2\bar{b}} \times \frac{\ln(1+\delta) + \sqrt{2\ln(1+\delta)}}{\ln(1+\delta)} \right) n + 1.$$

Proof. In this case, we can write

$$\begin{aligned} n \ln(2) &\geq \ln \prod_{i=1}^{d} \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \\ &= x_\varepsilon(\beta, L) - \frac{1}{2\bar{\beta}+1} \ln(1 + x_\varepsilon(\beta, L)). \end{aligned}$$

Thus, we have:

$$x_\varepsilon(\beta, L) \leq n\ln(2) + \ln(1 + x_\varepsilon(\beta, L)).$$

Now, let us remark that, if x is such that $x \leq A + \ln(1+x)$ for a given constant $A > 0$, then $x \leq A + \sqrt{2A}$. Thus, we have

$$\begin{aligned} x_\varepsilon(\beta, L) &\leq n\ln(2) + \sqrt{2n\ln(2)} \\ &\leq n\left(\ln(2) + \sqrt{2\ln(2)}\right). \end{aligned}$$

Now, let us write

$$\begin{aligned} \ln\frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} &= \left(\frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i} - \frac{2\bar{b}}{2\bar{b}+1}\frac{1}{b_i}\right)\ln\frac{l_*}{\|K\|\varepsilon} \\ &\quad + \frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i}\ln\frac{L}{l_*} \\ &\quad - \frac{\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i}\ln(1 + x_\varepsilon(\beta, L)). \end{aligned}$$

Let us estimate this quatity.

Upper bound. First, using the fact that $\bar{\beta} \leq \beta_i$ for all i, we obtain:

$$\ln\frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \leq \left(\frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i} - \frac{2\bar{b}}{2\bar{b}+1}\frac{1}{b_i}\right)\ln\frac{l_*}{\|K\|\varepsilon} + \frac{2}{2\bar{\beta}+1}\ln\frac{L}{l_*}.$$

On the other hand, it is easy to prove that:

$$\frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i} - \frac{2\bar{b}}{2\bar{b}+1}\frac{1}{b_i} \leq \frac{2\bar{b}+1}{2\bar{b}}\frac{4(\bar{b}-\bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)}.$$

Indeed, let us write:

$$\begin{aligned} \frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i} - \frac{2\bar{b}}{2\bar{b}+1}\frac{1}{b_i} &= \frac{2\bar{\beta}}{2\bar{\beta}+1}\left(\frac{1}{\beta_i} - \frac{1}{b_i}\right) - \frac{1}{b_i}\left(\frac{2\bar{b}}{2\bar{b}+1} - \frac{2\bar{\beta}}{2\bar{\beta}+1}\right) \\ &\leq \frac{2\bar{\beta}}{2\bar{\beta}+1}\left(\frac{1}{\beta_i} - \frac{1}{b_i}\right) \\ &\leq \frac{2\bar{\beta}}{2\bar{\beta}+1}\left(\frac{1}{\bar{\beta}} - \frac{1}{\bar{b}}\right) \\ &\leq \frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{\bar{b}-\bar{\beta}}{\bar{b}\bar{\beta}} \\ &= \frac{2\bar{b}+1}{2\bar{b}}\frac{4(\bar{b}-\bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)}. \end{aligned}$$

Finally, we obtain:

$$\ln \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \leq \frac{2\bar{b}+1}{2\bar{b}} x_\varepsilon(\beta, L),$$

and thus

$$\begin{aligned} k_i(\beta, L, \varepsilon) &\leq \frac{1}{\ln(2)} \ln \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \\ &\leq \frac{2\bar{b}+1}{2\bar{b}} \times \frac{x_\varepsilon(\beta, L)}{\ln(2)} \\ &\leq \left(\frac{2\bar{b}+1}{2\bar{b}} \times \frac{\ln(2) + \sqrt{2\ln(2)}}{\ln(2)} \right) n. \end{aligned}$$

Lower bound. First, let us supose that the following fact is proved:

$$\frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i} - \frac{2\bar{b}}{2\bar{b}+1}\frac{1}{b_i} \geq -\frac{1}{2\bar{b}} \frac{4(\bar{b}-\bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)}. \tag{A.1}$$

Then, we obtain

$$\ln \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \geq -\frac{1}{2\bar{b}} \frac{4(\bar{b}-\bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)} \ln \frac{l_*}{\|K\|\varepsilon} - \frac{\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i} \ln(1 + x_\varepsilon(\beta, L)).$$

Using the inequality $x_\varepsilon(\beta, L) \geq \ln l_*/\|K\|\varepsilon$, it follows

$$\ln \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \geq -\frac{1}{2\bar{b}} \left(x_\varepsilon(\beta, L) + \frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{\bar{b}}{\beta_i} \ln(1 + x_\varepsilon(\beta, L)) \right).$$

And, then

$$\ln \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \geq -\frac{2\bar{b}+1}{2\bar{b}} x_\varepsilon(\beta, L).$$

Finally:

$$\begin{aligned} k_i(\beta, L, \varepsilon) + 1 &\geq \frac{1}{\ln(2)} \ln \frac{h_i^*(\varepsilon)}{h_i(\beta, L, \varepsilon)} \\ &\geq -\left(\frac{2\bar{b}+1}{2\bar{b}} \times \frac{\ln(2) + \sqrt{2\ln(2)}}{\ln(2)} \right) n. \end{aligned}$$

To end the proof of this lema, we have to prove inequality (A.1) i.e.

$$(*) = \left(\frac{2\bar{\beta}}{2\bar{\beta}+1}\frac{1}{\beta_i} - \frac{2\bar{b}}{2\bar{b}+1}\frac{1}{b_i} \right) / \left(\frac{4(\bar{b}-\bar{\beta})}{(2\bar{b}+1)(2\bar{\beta}+1)} \right) \geq -\frac{1}{2\bar{b}}.$$

But,

$$\begin{aligned}(*) &= \frac{2\bar{\beta}(2\bar{b}+1)b_i - 2\bar{b}(2\bar{\beta}+1)\beta_i}{4b_i\beta_i(\bar{b}-\bar{\beta})} \\ &= \frac{2\bar{\beta}(2\bar{b}+1)(b_i-\beta_i) - 2\beta_i(\bar{b}-\bar{\beta})}{4b_i\beta_i(\bar{b}-\bar{\beta})} \\ &\geq -\frac{1}{2b_i}.\end{aligned}$$

A.3 Proof of lemma 4

Let us introduce a new notation. For all $i \in [\![1; d]\!]$, x and y in $\mathbf{R}^d$, let us denote:

$$[x, y]^{(i)} = (x_1, \dots, x_{i-1}, 0, y_{i+1}, \dots, y_d).$$

Let us fix k and l in $\mathcal{Z}_\varepsilon$. We are interrested in the following quatity:

$$b_{k\wedge l}(t) - b_l(t) = \int_{\mathbf{R}^d} K(u) \left(f(t - h^{(k\wedge l)}.u) - f(t - h^{(l)}.u) \right) du.$$

Let us consider the set $J \subset [\![1; d]\!]$ defined by:

$$J = \left\{ i \in [\![1; d]\!]; h_i^{(k)} > h_i^{(l)} \right\}.$$

If $i \in J^c$ then $h_i^{(k\wedge l)} = h_i^{(l)}$. Thus, we denote:

$$\eta_i = \begin{cases} h_i^{(k\wedge l)} = h_i^{(l)} \text{ if } i \in J^c \\ 0 \text{ if } i \in J. \end{cases}$$

Using these notations, we obtain:

$$\begin{aligned} b_{k\wedge l}(t) - b_l(t) &= \int_{\mathbf{R}^d} K(u) \left(f(t - h^{(k\wedge l)}.u) - f(t - \eta.u) \right) du \\ &\quad + \int_{\mathbf{R}^d} K(u) \left(f(t - \eta.u) - f(t - h^{(l)}.u) \right) du. \end{aligned}$$

Thus, it is enough to study quantities of the following form:

$$\int_{\mathbf{R}^d} K(u) \left(f(t - h.u) - f(t - \eta.u) \right) du$$

where $h = h^{(k\wedge l)}$ else $h = h^{(l)}$. In both case we have $h_i = \eta_i$ if $i \in J^c$ and $h_i \leq h_i^{(k)}$ if $i \in J$. Here and later we will consider a such bandwidth h.

It is easy to rewrite the following quatity

$$(*) = f(t - h.u) - f(t - \eta.u),$$

using a telescopic sum. We obtain:

$$(*) \;=\; \sum_{i=1}^{d} f_i(-h_i u_i | t - [\eta, h]^{(i)}.u) - f_i(-\eta_i u_i | t - [\eta, h]^{(i)}.u).$$

As $h_i u_i = \eta_i u_i$ if $i \in J^c$, we deduce that indexes belonging to J^c do not contribute to the sum. Finally:

$$(*) = \sum_{i \in J} f_i(-h_i u_i | t - [\eta, h]^{(i)}.u) - f_i(0 | t - [\eta, h]^{(i)}.u).$$

Now, using that f belongs to the anisotropic class $H(\beta, L)$ it is easy to develop the quantity $(*)$ using a Taylor's formula. If we denote $m_i = \lfloor \beta_i \rfloor$, we obtain:

$$\begin{aligned}(*) \;=\; & \sum_{i \in J} \sum_{n=1}^{m_i} f_i^{(n)}(0 | t - [\eta, h]^{(i)}.u) \frac{(-h_i u_i)^n}{n!} \\ & + \sum_{i \in J} \frac{(-h_i u_i)^{m_i}}{m_i!} \left(f_i^{(m_i)}(\theta_i | t - [\eta, h]^{(i)}.u) - f_i^{(m_i)}(0 | t - [\eta, h]^{(i)}.u) \right),\end{aligned}$$

where $|\theta_i| \leq h_i |u_i|$.

If we remark that $t - [\eta, h]^{(i)}.u$ does not depend on u_i, using hypothesis $(K4)$ on K and Fubini's theorem, we obtain that, for all $i \in J$ and $n \in [\![1; m_i]\!]$, we have:

$$\int_{\mathbf{R}^d} K(u) f_i^{(n)}(0 | t - [\eta, h]^{(i)}.u) \frac{(-h_i u_i)^n}{n!} du = 0.$$

Moreover it is easy to obtain that if $i \in J$, then:

$$\begin{aligned}\left| f_i^{(m_i)}(\theta_i | t - [\eta, h]^{(i)}.u) - f_i^{(m_i)}(0 | t - [\eta, h]^{(i)}.u) \right| \;&\leq\; L |\theta_i|^{\beta_i - m_i} \\ &\leq\; L h_i^{\beta_i - m_i} |u_i|^{\beta_i - m_i}.\end{aligned}$$

Then, we can deduce that:

$$\begin{aligned}\left| \int_{\mathbf{R}^d} K(u) \left(f(t - h.u) - f(t - \eta.u) \right) du \right| \;&\leq\; L \sum_{i \in J} \left(\int_{\mathbf{R}^d} |K(u)| \frac{|u_i|^{\beta_i}}{m_i!} du \right) h_i^{\beta_i} \\ &\leq\; L \sum_{i=1}^{d} \lambda_i(\beta) h_i^{\beta_i}.\end{aligned}$$

Lemma follows.

A.4 Proof of lemma 5

First of all, let us remark that:

$$\forall i, \forall (\beta, L), \quad 1 \leq \frac{h_i(\beta, L, \varepsilon)}{h_i^{(k(\beta,L,\varepsilon))}} \leq 2.$$

Let us calculate:

$$\begin{aligned} B^{\beta,L}(k(\beta, L, \varepsilon)) &= b^{\beta,L}(h^{(k(\beta,L,\varepsilon))}) \\ &= L \sum_{i=1}^{d} \lambda_i(\beta) \left(h_i^{(k(\beta,L,\varepsilon))}\right)^{\beta_i} \\ &= L \sum_{i=1}^{d} \lambda_i(\beta) \left(h_i(\beta, L, \varepsilon)\right)^{\beta_i} \left(\frac{h_i^{(k(\beta,L,\varepsilon))}}{h_i(\beta, L, \varepsilon)}\right)^{\beta_i} \\ &\leq b^{\beta,L}(h(\beta, L, \varepsilon)). \end{aligned}$$

On the other hand, it is easy to prove that:

$$b^{\beta,L}(h(\beta, L, \varepsilon)) = \left(\sum_{i=1}^{d} \lambda_i(\beta)\right) s_\varepsilon(h(\beta, L, \varepsilon)).$$

Thus, we obtain:

$$B^{\beta,L}(k(\beta, L, \varepsilon)) \leq (d\lambda^*) s_\varepsilon(h(\beta, L, \varepsilon)).$$

Now, we focus our attention on $s_\varepsilon(h(\beta, L, \varepsilon))$. First, it is easy to prove that:

$$s_\varepsilon(h(\beta, L, \varepsilon)) \leq \frac{\rho_\varepsilon(\beta, L)}{\sqrt{1 + |k(\beta, L, \varepsilon)| \ln 2}} S_\varepsilon(k(\beta, L, \varepsilon)).$$

Next, let us prove that:

$$\frac{\rho_\varepsilon^2(\beta, L)}{1 + |k(\beta, L, \varepsilon)| \ln 2} = \frac{1 + x_\varepsilon(\beta, L)}{1 + |k(\beta, L, \varepsilon)| \ln 2} \leq 2 \vee \frac{\bar{b} + 1}{\bar{b}}.$$

It is known that:

$$1 + |k(\beta, L, \varepsilon)| \ln 2 \geq 1 + x_\varepsilon(\beta, L) - \frac{1}{2\bar{\beta} + 1} \ln(1 + x_\varepsilon(\beta, L)), \tag{A.2}$$

thus, we obtain that:

$$x_\varepsilon(\beta, L) \leq |k(\beta, L, \varepsilon)| \ln 2 + \frac{1}{2\bar{\beta} + 1} \ln(1 + x_\varepsilon(\beta, L)).$$

This implies in particular that:

$$x_\varepsilon(\beta, L) \leq |k(\beta, L, \varepsilon)| \ln 2 + \sqrt{|k(\beta, L, \varepsilon)| \ln 2}.$$

If $|k(\beta, L, \varepsilon)| \geq 2$ (for example if $\bar{\beta} < \bar{b}/2$), we obtain:

$$x_\varepsilon(\beta, L) \leq 2|k(\beta, L, \varepsilon)| \ln 2,$$

which implies immediatly that:

$$\frac{1 + x_\varepsilon(\beta, L)}{1 + |k(\beta, L, \varepsilon)| \ln 2} \leq 2.$$

On the other hand, when $\bar{\beta} \geq \bar{b}/2$, we obtain from (A.2) that:

$$\begin{aligned} 1 + |k(\beta, L, \varepsilon)| \ln 2 &\geq 1 + x_\varepsilon(\beta, L) - \frac{1}{\bar{b} + 1} \ln(1 + x_\varepsilon(\beta, L)) \\ &\geq 1 + \frac{\bar{b}}{\bar{b} + 1} x_\varepsilon(\beta, L). \end{aligned}$$

Last inequality implies that:

$$\frac{1 + x_\varepsilon(\beta, L)}{1 + |k(\beta, L, \varepsilon)| \ln 2} \leq \frac{\bar{b} + 1}{\bar{b}}.$$

Lemma follows.

A.5 Proof of lemma 6

First, it is easy to prove that;

$$S_\varepsilon(k(\beta, L, \varepsilon)) \leq 2^{d/2} \sqrt{\frac{1 + |k(\beta, L, \varepsilon)| \ln 2}{1 + x_\varepsilon(\beta, L)}} s_\varepsilon(h(\beta, L, \varepsilon)).$$

Next, we know that:

$$1 + |k(\beta, L, \varepsilon)| \ln 2 = 1 + x_\varepsilon(\beta, L) - \frac{1}{2\bar{\beta}} \ln(1 + x_\varepsilon(\beta, L)) \leq 1 + x_\varepsilon(\beta, L).$$

Lemma is proved.

Appendix B

Partially Case

B.1 Proof of lemma 8

First of all, let us prove that $a < b_1$. In fact:

$$a < b_1 \iff \frac{1}{b_1} < \frac{1}{\gamma} - \frac{1}{\delta}.$$

But it is clear that

$$\frac{1}{b_1} + \frac{1}{\delta} = \frac{1}{\overline{b}} < \frac{1}{\gamma}.$$

Result follows.

Let us fix $\beta \in \{\beta^{(k)}\}_k$.

Step 1. Let us calculate:

$$\begin{aligned} \sum_{i=1}^{d} \frac{1}{\beta_i} &= \frac{1}{\beta_1} + \sum_{i=2}^{d} \frac{\delta}{b_i} \left(\frac{1}{\gamma} - \frac{1}{\beta_1} \right) \\ &= \frac{1}{\gamma}. \end{aligned}$$

Step 2. Let us prove that, for all i, $\beta_i > 0$. First, we have $\beta_1 > a > 0$. Next, for $i \geq 2$, $\beta_i > 0$ if $1/\gamma > 1\beta_1$. But clearly we have $\beta_1 > a > \gamma$. Result follows.

Step 3. Let us prove that, for all i, $\beta_i \leq b_i$. This inequality is equivalent to:

$$\delta\left(\frac{1}{\gamma} - \frac{1}{\beta_1}\right) \geq 1,$$

i.e. $\beta_1 \geq a$. Finally, $\beta \in \mathcal{B}(\gamma)$. □

B.2 Proof of lemma 9

First, let us remark that:

$$\mathbf{P}_0[|Z_\varepsilon - 1| > \alpha] \leq \alpha^{-2}\mathbf{E}_0\big[(Z_\varepsilon - 1)^2\big]$$

and, if $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbf{L}^2$,

$$\mathbf{E}_0\big[(Z_\varepsilon - 1)^2\big] = \frac{1}{n_\varepsilon^2}\sum_{k,l=1}^{n_\varepsilon} \exp\left(\frac{\langle f_k, f_l\rangle}{\varepsilon^2}\right) - 1.$$

It is enough to prove the following assertions:

$$\frac{1}{n_\varepsilon^2}\sum_{k=1}^{n_\varepsilon} \exp\left(\frac{\|f_k\|}{\varepsilon^2}\right) \underset{\varepsilon\to 0}{\longrightarrow} 0, \tag{B.1}$$

and

$$\limsup_{\varepsilon\to 0} \frac{1}{n_\varepsilon^2}\sum_{k\neq l}^{n_\varepsilon} \exp\left(\frac{\langle f_k, f_l\rangle}{\varepsilon^2}\right) \leq 1 \tag{B.2}$$

First, let us prove Equation (B.1).

Let us calculate $\|f_k\|^2$ for all k. We have:

$$\|f_k\|^2 = \|g\|^2\varkappa^2\varepsilon^2 \ln\ln\frac{1}{\varepsilon} = 2\|g\|^2\varkappa^2\varepsilon^2\ln n_\varepsilon.$$

Thus, we obtain:

$$\frac{1}{n_\varepsilon^2}\sum_{k=1}^{n_\varepsilon} \exp\left(\frac{\|f_k\|}{\varepsilon^2}\right) = n_\varepsilon^{2\|g\|^2\varkappa^2 - 1}.$$

Thus, the choice of $\varkappa$ implies the result because $2\|g\|^2\varkappa^2 - 1 < 0$.

Now, let us prove Equation (B.2).

Let us fix $1 \leq k < l \leq n_\varepsilon$. By an easy computation we obtain:

$$\langle f_k, f_l \rangle \leq \varkappa^{\frac{4\gamma}{2\gamma+1}} \eta_\varepsilon^2(\gamma) \|g\|_\infty^2 \mathrm{Vol}(C_k \cap C_l),$$

where Vol is the standard volume in $\mathbf{R}^d$ and C_k denotes the support of f_k:

$$C_k = \prod_{i=1}^{d} [-a h_i^{(k)}; a h_i^{(k)}].$$

Clearly, $h_1^{(k)} < h_1^{(l)}$ and, for any $i \geq 2$, we have $h_i^{(k)} > h_i^{(l)}$. Thus,w e can conclude that:

$$\mathrm{Vol}(C_k \cap C_l) = (2a)^d \frac{h_1^{(k)}}{h_1^{(l)}} \left(\prod_{i=1}^{d} h_i^{(l)} \right) \leq (2a)^d \frac{h_1^{(k)}}{h_1^{(k+1)}} \left(\prod_{i=1}^{d} h_i^{(l)} \right).$$

Let us calculate $h_1^{(k)}/h_1^{(k+1)}$:

$$\begin{aligned} \frac{h_1^{(k)}}{h_1^{(k+1)}} &= \left(\varkappa^{\frac{2\gamma}{2\gamma+1}} \eta_\varepsilon(\gamma) \right)^{1/\beta_1^{(k)} - 1/\beta_1^{(k+1)}} \\ &= \left(\varkappa^{\frac{2\gamma}{2\gamma+1}} \eta_\varepsilon(\gamma) \right)^{\frac{1/n_\varepsilon}{\beta_1^{(k)} \beta_1^{(k+1)}}} \\ &\leq \left(\varkappa^{\frac{2\gamma}{2\gamma+1}} \eta_\varepsilon(\gamma) \right)^{\frac{1}{b_1^2 n_\varepsilon}}. \end{aligned}$$

Moreover, let us remark that:

$$\prod_{i=1}^{d} h_i^{(l)} = \varkappa^{\frac{2}{2\gamma+1}} \eta_\varepsilon^{1/\gamma}(\gamma).$$

Then, by an easy computation, we deduce that:

$$\langle f_k, f_l \rangle \leq (2a)^d (\varkappa^{1+\frac{\Gamma}{n_\varepsilon}} \|g\|_\infty)^2 \left(\eta_\varepsilon(\gamma) \right)^{\frac{2\gamma+1}{\gamma}\left(1+\frac{\Gamma}{n_\varepsilon}\right)},$$

where

$$\Gamma = \frac{\gamma}{b_1^2(2\gamma+1)}.$$

Let us recall that $\eta_\varepsilon(\gamma) = (\varepsilon \sqrt{\ln \ln 1/\varepsilon})^{2\gamma/(2\gamma+1)}$. Thus we obtain:

$$\frac{\langle f_k, f_l \rangle}{\varepsilon^2} \leq (2a)^d (\varkappa^{1+\frac{\Gamma}{n_\varepsilon}} \|g\|_\infty)^2 \mathcal{M}_\varepsilon,$$

where

$$\mathcal{M}_\varepsilon = \left(\ln\ln\frac{1}{\varepsilon}\right)\left(\varepsilon^2\ln\ln\frac{1}{\varepsilon}\right)^{\frac{\Gamma}{n_\varepsilon}}$$

teds to 0 when ε tends to 0 (it is easy to see that $\ln\mathcal{M}_\varepsilon \to -\infty$).

Now, let us back to Equation (B.2):

$$\frac{1}{n_\varepsilon^2}\sum_{k\neq l}^{n_\varepsilon}\exp\left(\frac{\langle f_k, f_l\rangle}{\varepsilon^2}\right) \leq \frac{n_\varepsilon - 1}{n_\varepsilon}\exp\left((2a)^d(\varkappa^{1+\frac{\Gamma}{n_\varepsilon}}\|g\|_\infty)^2\mathcal{M}_\varepsilon\right) \underset{\varepsilon\to 0}{\longrightarrow} 1.$$

And Lemma is proved. □

Bibliographie

[1] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413, 1999.

[2] E. N. Belitser and B. Y. Levit. On minimax filtering over ellipsoids. *Math. Methods Statist.*, 4(3) :259–273, 1995.

[3] E. N. Belitser and B. Y. Levit. Asymptotically minimax nonparametric regression in L_2. *Statistics*, 28(2) :105–122, 1996.

[4] Karine Bertin. Asymptotically exact minimax estimation in sup-norm for anisotropic Hölder classes. *Bernoulli*, 10(5) :873–888, 2004.

[5] Karine Bertin. Minimax exact constant in sup-norm for nonparametric regression with random design. *J. Statist. Plann. Inference*, 123(2) :225–242, 2004.

[6] Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2) :181–237, 1983.

[7] Lucien Birgé. Nonasymptotic minimax risk for Hellinger balls. *Probab. Math. Statist.*, 5(1) :21–29, 1985.

[8] Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[9] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) :203–268, 2001.

[10] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3) :843–874, 2002. Dedicated to the memory of Lucien Le Cam.

[11] L. Cavalier, Y. Golubev, O. Lepski, and A. Tsybakov. Block thresholding and sharp adaptive estimation in severely ill-posed inverse problems. *Teor. Veroyatnost. i Primenen.*, 48(3) :534–556, 2003.

[12] L. Cavalier and A. B. Tsybakov. Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist.*, 10(3) :247–282, 2001. Meeting on Mathematical Statistics (Marseille, 2000).

[13] Albert Cohen, Ronald DeVore, Gerard Kerkyacharian, and Dominique Picard. Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 11(2) :167–191, 2001.

[14] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994.

[15] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432) :1200–1224, 1995.

[16] David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage : asymptopia ? *J. Roy. Statist. Soc. Ser. B*, 57(2) :301–369, 1995. With discussion and a reply by the authors.

[17] David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2) :508–539, 1996.

[18] S. Yu. Efroĭmovich and M. S. Pinsker. Estimation of square-integrable probability density of a random variable. *Problemy Peredachi Informatsii*, 18(3) :19–38, 1982.

[19] Sam Efromovich. *Nonparametric curve estimation*. Springer Series in Statistics. Springer-Verlag, New York, 1999. Methods, theory, and applications.

[20] V. N. Gabušin. Best approximation of functionals on certain sets. *Mat. Zametki*, 8 :551–562, 1970.

[21] A. Goldenshluger and A. Nemirovski. On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.*, 6(2) :135–170, 1997.

[22] G. K. Golubev. Adaptive asymptotically minimax estimates for smooth signals. *Problemy Peredachi Informatsii*, 23(1) :57–67, 1987.

[23] Y. Golubev and B. Levit. An oracle approach to adaptive estimation of linear functionals in a Gaussian model. *Math. Methods Statist.*, 13(4) :392–408 (2005), 2004.

[24] Yu. Golubev, O. Lepski, and B. Levit. On adaptive estimation for the sup-norm losses. *Math. Methods Statist.*, 10(1) :23–37, 2001.

[25] Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

[26] R. Z. Hasminskii and I. A. Ibragimov. On asymptotic efficiency in the presence of an infinite-dimensional nuisance parameter. In *Probability theory and mathematical statistics (Tbilisi, 1982)*, volume 1021 of *Lecture Notes in Math.*, pages 195–229. Springer, Berlin, 1983.

[27] I. A. Ibragimov and R. Z. Hasminski. On the estimation of a signal, its derivatives and the maximum point for Gaussian observations. *Teor. Veroyatnost. i Primenen.*, 25(4) :718–733, 1980.

[28] I. A. Ibragimov and R. Z. Hasminski. Some problems of nonparametric estimation of the value of a linear functional. *Dokl. Akad. Nauk SSSR*, 256(4) :781–783, 1981.

[29] I. A. Ibragimov and R. Z. Hasminski. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York, 1981. Asymptotic theory, Translated from the Russian by Samuel Kotz.

[30] Ildar Abdulovic Ibragimov and Rafail Zalmanovic Hasminski. Asymptotic properties of some nonparametric estimates in Gaussian white noise. In *Third International Summer School on Probability Theory and Mathematical Statistics (Varna, 1978) (Bulgarian)*, pages 29–64. B lgar. Akad. Nauk, Sofia, 1980.

[31] Gérard Kerkyacharian, Oleg Lepski, and Dominique Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields*, 121(2) :137–170, 2001.

[32] Gérard Kerkyacharian and Dominique Picard. Estimating nonquadratic functionals of a density using Haar wavelets. *Ann. Statist.*, 24(2) :485–507, 1996.

[33] Gérard Kerkyacharian and Dominique Picard. Minimax or maxisets ? *Bernoulli*, 8(2) :219–253, 2002.

[34] Gérard Kerkyacharian, Dominique Picard, and Karine Tribouley. L^p adaptive density estimation. *Bernoulli*, 2(3) :229–247, 1996.

[35] A. P. Korostelev. An asymptotically minimax regression estimator in the uniform norm up to a constant. *Teor. Veroyatnost. i Primenen.*, 38(4) :875–882, 1993.

[36] A. P. Korostelev and A. B. Tsybakov. Asymptotically minimax image reconstruction problems. In *Topics in nonparametric estimation*, volume 12 of *Adv. Soviet Math.*, pages 45–86. Amer. Math. Soc., Providence, RI, 1992.

[37] Alexander Korostelev and Michael Nussbaum. The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. *Bernoulli*, 5(6) :1099–1118, 1999.

[38] O. V. Lepski. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3) :459–470, 1990.

[39] O. V. Lepski. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4) :645–659, 1991.

[40] O. V. Lepski and B. Y. Levit. Adaptive minimax estimation of infinitely differentiable functions. *Math. Methods Statist.*, 7(2) :123–156, 1998.

[41] O. V. Lepski and B. Y. Levit. Adaptive nonparametric estimation of smooth multivariate functions. *Math. Methods Statist.*, 8(3) :344–370, 1999.

[42] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness : an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3) :929–947, 1997.

[43] O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.*, 25(6) :2512–2546, 1997.

[44] O. V. Lepski and A. B. Tsybakov. Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probab. Theory Related Fields*, 117(1) :17–48, 2000.

[45] O. V. Lepskiĭ. On problems of adaptive estimation in white Gaussian noise. In *Topics in nonparametric estimation*, volume 12 of *Adv. Soviet Math.*, pages 87–106. Amer. Math. Soc., Providence, RI, 1992.

[46] E. Mammen. On qualitative smoothness of kernel density estimates. *Statistics*, 26(3) :253–267, 1995.

[47] Enno Mammen. A short note on optimal bandwidth selection for kernel estimators. *Statist. Probab. Lett.*, 9(1) :23–25, 1990.

[48] Michael Nussbaum. Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, 24(6) :2399–2430, 1996.

[49] Valentin V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.

[50] Dominique Picard and Karine Tribouley. Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.*, 28(1) :298–335, 2000.

[51] Vincent Rivoirard. Maxisets for linear procedures. *Statist. Probab. Lett.*, 67(3) :267–275, 2004.

[52] Vincent Rivoirard. Thresholding procedure with priors based on Pareto distributions. *Test*, 13(1) :213–246, 2004.

[53] A. V. Skorohod. *Integration in Hilbert space*. Springer-Verlag, New York, 1974. Translated from the Russian by Kenneth Wickwire, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 79.

[54] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6) :1348–1360, 1980.

[55] Charles J. Stone. Admissibility and local asymptotic admissibility of procedures which combine estimation and model selection. In *Statistical decision theory and related topics, III, Vol. 2 (West Lafayette, Ind., 1981)*, pages 317–333. Academic Press, New York, 1982.

[56] Charles J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12(4) :1285–1297, 1984.

[57] Charles J. Stone. Additive regression and other nonparametric models. *Ann. Statist.*, 13(2) :689–705, 1985.

[58] Hans Triebel. *Theory of function spaces. II*, volume 84 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel, 1992.

[59] A. B. Tsybakov. Asymptotically efficient estimation of a signal in L_2 under general loss functions. *Problemy Peredachi Informatsii*, 33(1) :94–106, 1997.

[60] A. B. Tsybakov. Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.*, 26(6) :2420–2469, 1998.

[61] Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.

Sur l'estimation adaptatives de fonctions anistropes

Résumé. Cette thèse est consacrée à l'étude de problèmes statistiques d'estimation non paramétrique. Un signal bruité multidimensionnel est observé (par exemple une image dans le cas de la dimension deux) et nous nous fixons l'objectif de le reconstruire *au mieux*.

Pour réaliser ce but, nous nous plaçons dans le cadre de la théorie adaptative au sens minimax : nous cherchons un seul estimateur qui atteint simultanément sur chaque espace fonctionnel d'une collection la « meilleure vitesse possible ».

Nous donnons un nouveau critère pour choisir une famille de normalisations optimale. Ce critère est plus sophistiqué que ceux introduits par Lepski (1991) puis Tsybakov (1998) et est mieux adapté au cas multidimensionnel.

Ensuite, nous donnons deux résultats adaptatifs (en estimation ponctuelle) par rapport à deux collections différentes d'espaces de Hölder anisotropes. Dans les deux cas, nous construisons des procédures (basées sur la comparaison d'estimateurs à noyau pour choisir, en fonction des observations, le meilleur d'eux) dont nous prouvons qu'elles sont optimales en un certain sens.

On the adaptive estimation of anisotropic functions

Abstract. This thesis is devoted to the study of statistical problems of non parametrical estimation. A noisy multidimensionnal signal is observed (for example an image if the dimension is equal to two) and our goal is to reconstruct it *as best as possible*.

In order to achieve this goal, we consider the well known theory of adaptation on a minimax sense : we want to construct a single estimator which achieves on each fuctionnal space of a given collection the « best possible rate ».

We introduce a new criterion in order to chose an optimal family of normalizations. This criterion is more sophisticated than criteria given by Lepski (1991) and Tsybakov (1998) and well adapted to multidimensionnal case.

Then, we prove two results of adaptation with respect to different collections of anisotropic Hölder spaces. In each case, we construct a procedure (based on comparison of kernel estimators in order to chose, according to the observations, the best among a family) which is proved to be optimal in a given sense.